Pronunciation Variation Modeling for Dutch Automatic Speech Recognition

Image and design - Andrew Guy Smith
Printed and bound by Ponsen & Looijen bv, Wageningen

ISBN: 90-9015608-9

Pronunciation Variation Modeling for Dutch Automatic Speech Recognition

Een wetenschappelijke proeve op het gebied van de Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen in het
openbaar te verdedigen op dinsdag 15 april 2002
des namiddags om 3:30 uur precies

door

Mirjam Wester

geboren op 24 augustus 1971
te Delft

Promotor:	Prof. dr. L.W.J. Boves
Co-promotor:	Dr. W.A.J. Strik
Manuscriptcommissie:	Prof. dr. R. van Hout
	Prof. dr. ir. J.-P. Martens (Universiteit Gent)
	Dr. S. Greenberg (International Computer Science Institute)

Acknowledgements

This thesis reports on research that was carried out at the University of Nijmegen, the Netherlands and at the International Computer Science Institute (ICSI) in Berkeley, U.S.A. As with almost all research carried out in the field of speech technology, this was by no means a solo-project, and it would not have been possible without the help, expertise and guidance of other people. I would like to use this opportunity to acknowledge and thank those people who contributed to the writing and realization of this thesis.

First of all, I would like to thank Helmer Strik for supervising this project. I especially appreciate the fact that he gave me the freedom and space to do it my way, and for supporting my choice to carry out further research at ICSI. Without his advice, comments and supervision I am sure the project would have stranded along the way. In addition to the supervision Helmer gave me, I was in the fortunate position of having Lou Boves as my promotor. I would like to thank Lou for all the useful feedback he gave at every stage of this PhD project and for the way he always managed to convince me that continuing with the thesis still made sense. I would like to express my gratitude to Judith Kessens for her close collaboration during the first years of the project and for being a good sounding board when I returned from ICSI. Thanks are due to Catia Cucchiarini for sharing her expertise, knowledge and writing skills. Each of the A^2RT members (past and present) also has to be thanked for generously sharing their knowledge and for all the time that went into reading numerous drafts. The group as a whole created a comfortable research environment, which contributed to my research.

Next, I would like to acknowledge and thank the following organizations: Netherlands Organization for Scientific Research (NWO), Shell Nederland B.V., International Speech Communication Association (ISCA) and the European Union (TMR grant) for their financial support. Without this support I would not have been able to attend the various workshops and international conferences which I have had the privilege of attending in the last five years – the real perks of the job.

I am very grateful for the opportunity which I had to visit ICSI for a year as a guest researcher, which was made possible by the “Frye stipendium”, an NWO travel grant and the U.S. Department of Defense. I would like to thank everybody at ICSI for making my time in Berkeley a valuable and enjoyable learning experience. In particular, I would like to extend my gratitude to Eric Fosler-Lussier with whom I had the priveledge of collaborating during the first 6 months of my stay. I would further like to thank Steve Greenberg for inviting me to

extend my stay at ICSI, and for sharing his expertise, knowledge and time. It was a pleasure to work with both him and Shawn Chang. In addition, I would like to thank Lila Finhill for making me feel welcome at ICSI and for always being ready for a chat.

Andrew Smith is thanked for using his talents to create a wonderful cover that livens up a piece of work which from a design perspective would otherwise have been fairly bleak. Febe de Wet is thanked for proofreading the thesis and all the other times she quickly read and commented on other draft papers and articles. I would also like to thank her for making life in Nijmegen more enjoyable and being my “paranimf”.

Thanks are due to my Dad and my eldest brother Flip, after all they are the reason I got into all of this! I am extremely grateful to my parents and siblings for creating the intellectually stimulating environment I grew up in and for supporting me. Finally, I would like to thank Angus for all his support and for sticking with me through a collapsed lung and all.

Contents

Part I

Introductory review

1.1 Introduction to pronunciation variation

Speech is variable. The way in which a sound, word or sequence of words is pronounced can be different every time it is produced (Strik and Cucchiarini 1999). This pronunciation variation can be the result of:

- Intra-speaker variability: the variation in pronunciation for one and the same speaker.
- Inter-speaker variability: the variation among different speakers. The variation can be due to factors such as vocal tract differences, age, gender, regional accent, dialect, voice quality etc. (Laver 1968; Biemans 2000).

There are numerous factors that influence the degree of intra-speaker pronunciation variation that is encountered in speech. These include:

- Speaking style, also referred to as stylistic variation: this type of variation depends on whether the speech is scripted, planned or spontaneous (Weintraub et al. 1996).
- Speaking rate: it has been shown that speaking rate can have a dramatic impact on the degree of variation in pronunciation (Greenberg and Fosler-Lussier 2000).
- Coarticulation: the overlapping of adjacent articulations affects the way words are pronounced (Ladefoged 1975), and variation in the degree of coarticulation causes pronunciation variation.
- Suprasegmental features: for instance, word stress, sentence stress, intonation, frequency of occurrence of a word, position of a word in a sentence, and position of a consonant or vowel within a syllable all affect the pronunciation of words (Ladefoged 1975; Greenberg and Chang 2000).
- State of health of the speaker: factors such as whether the speaker has a cold or is fatigued influences how the words are pronounced.
- Emotional state of the speaker: whether the speaker is happy, sad, or excited (Polzin and Waibel 1998), but also factors such as the speaker's attitude towards the topic under discussion has an effect on the pronunciation.
- External conditions: for instance noise, which causes speakers to modify their speech: the Lombard effect (Junqua 1993).
- The interlocutor: people speak in different ways depending on who they are speaking to; for instance a child or an adult. Another relevant example is the situation where the interlocutor is a computer system. Under such circumstances a speaker may be influenced in the way he/she pronounces the words, as well.

These sources of variation all contribute to the fact that a word is never pronounced in exactly the same way by the same or different speakers. This is referred to as pronunciation variation.

The objective of automatic speech recognition (ASR) is to recognize what a person has said, i.e., to derive the string of spoken words from an acoustic signal. Due to the above described variation this objective becomes more difficult to achieve, as the pronunciation variation may lead to recognition errors. Therefore, avenues are sought to model pronunciation variation. The type of pronunciation variation that is focussed on in this thesis is variation that becomes apparent in a careful phonetic transcription of speech, in the form of insertions, deletions or substitutions of phones relative to a single, normative ("canonical") transcription of the words.

This thesis consists of four articles (Part II), preceded by an introductory review (Part I). In the introductory review, the main themes in pronunciation modeling are discussed. In Section 1.2 a short introduction to ASR is given. This brief ASR introduction is intended to provide a framework for the discussion of the issues concerning pronunciation modeling. It is followed in Section 1.3 by arguments with regard to why pronunciation modeling is necessary for ASR. In Section 1.4, a number of issues that are relevant to pronunciation variation modeling are discussed. This is followed by a description of the speech material that was used in this thesis in Section 1.5. Summaries of the four articles are included in Section 1.6. A discussion of the shortcomings of previous and current approaches to the modeling of pronunciation variation is presented in Section 1.7, followed by the major conclusions of this work and future avenues worth exploring. Part II of this thesis consists of reprints of the four publications.

1.2 A brief introduction to ASR

In very general terms, the task of ASR is to derive a string of words from a stream of acoustic information. Figure 1.1 gives a schematic impression of the components that are involved in the speech recognition process. In this section, first the role of each of the components is discussed in more detail. This is followed by short descriptions of the two systems that were employed in this research: the Phicos system (Steinbiss et al. 1993) and the ICSI system (Bourlard and Morgan 1993). For more complete and detailed introductions to ASR, see Rabiner and Juang (1993) and Bourlard and Morgan (1993).

Feature extraction. In order to perform speech recognition, a recording of the speech signal is needed. It is the task of the feature extraction module (often referred to as the front-end) to convert the raw acoustic waveform into acoustic feature vectors. The objective of the front end/feature extraction module is to derive acoustic representations that are good at separating different classes of speech sounds and effective at suppressing irrelevant sources of variation. ASR typically uses features based on a short-term spectrum of speech. Feature vectors are computed using a local analysis window (termed a frame) of the order of 16-32 ms. Whatever the acoustic features are – e.g. MFCCs (Mel Frequency Cepstral Coefficients) or PLP features (Perceptual Linear

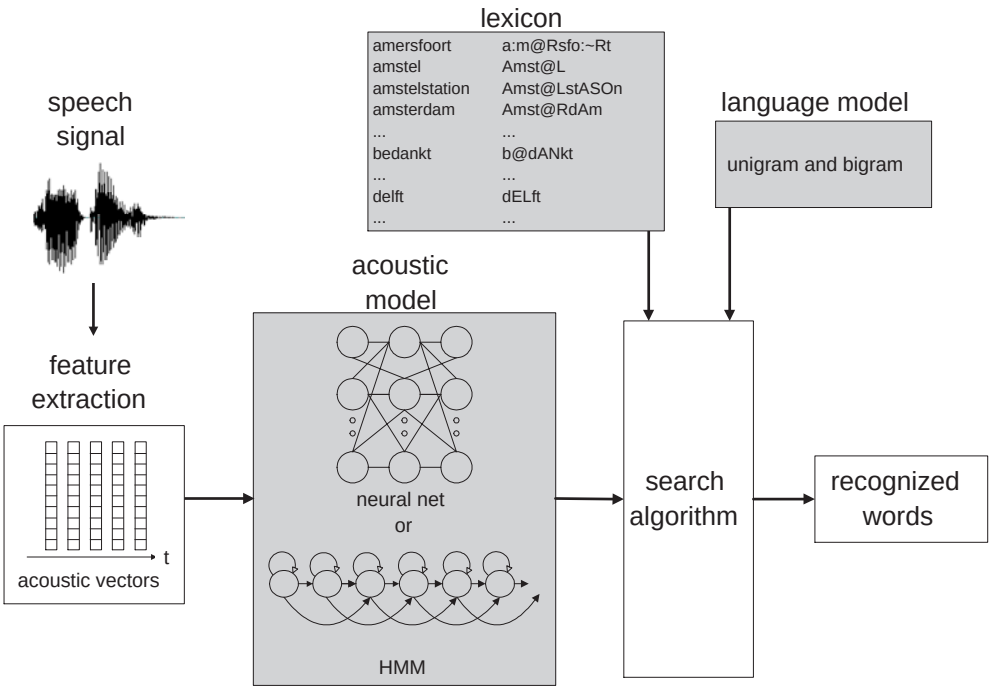


Figure 1.1: Overview of a speech recognizer, shaded areas indicate where pronunciation variation modeling is incorporated in the work presented in this thesis.

Prediction) – the feature extraction process converts the speech signal into a sequence of acoustic vectors, which can be symbolically represented as: $X = \{x_1, x_2, ..., x_T\}$, where T corresponds to the number of frames in the utterance.

Decoding. The speech recognition problem can then be formulated as:

$$\hat{W} = \underset{W \in \mathcal{W}}{\text{argmax}}\, P(W|X) \tag{1.1}$$

with $\mathcal{W}$ being the set of possible word sequences. Thus, the problem is to maximize over all possible word sequences W to obtain the highest probability P given the acoustics X.

Because of the extremely large number of possible word sequences in natural language, and the enormous range of variation in the acoustic signals that is produced when different speakers pronounce the "same" sequence of words, $P(W|X)$ cannot be computed directly. In order to deal with this problem, Bayes' rule is used to break up

this probability into components:

$$\hat{W} = \underset{W}{\text{argmax}} \frac{P(X|W)P(W)}{P(X)} \tag{1.2}$$

As the prior probability $P(X)$ in the denominator of Eq. 1.2 is constant over all W, Eq. 1.2 may be simplified to:

$$\hat{W} = \underset{W}{\text{argmax}} P(X|W)P(W) \tag{1.3}$$

Thus, the ASR system must model two probability distributions: $P(X|W)$ which is the posterior probability of the acoustics given a string of words, and is modeled by the acoustic models; and $P(W)$, the prior probability of a string of words, which is modeled by the language model. The set of possible words is defined in the lexicon. In the following paragraphs, the workings of the acoustic models, the lexicon, and the language model will be explained.

Acoustic model. The acoustic models are statistical models which capture the correspondence between a short sequence of acoustic vectors and an elementary unit of speech. The elementary units of speech that are most often used in ASR are phone(me)s. *Phonemes* are the minimal units of speech that are part of the sound system of a language, which serve to distinguish one word from another. Sounds which count as alternative ways of expressing one and the same phoneme are called *allophones*; in other words allophones are variants of one and the same phoneme. The term *phones* covers both phonemes and allophones.

The predominant approach to acoustic modeling in speech recognition is to use hidden Markov models (HMMs). An alternative to the standard HMM approach is a hybrid approach in which artificial neural networks (ANN) and HMMs are employed.

In order to recognize speech, the acoustic models must first be trained. During training, the parameters for the models are estimated from recorded speech material which has been orthographically transcribed (i.e., at word level). Additionally, a phonetic transcription of the words is needed. Transforming a word sequence to a phone sequence is accomplished by looking up the phonetic transcription for a word in the lexicon.

An HMM is a stochastic automaton, consisting of a collection of states connected by transitions (cf. Fig. 1.1). Two sets of probabilities are associated with each state: a transition probability, which gives the probability of taking the transition, and an output or emission probability density function, which specifies the probability of emitting each output symbol. An HMM is trained for each recognition unit (e.g. phones) defined in the system.

In a hybrid recognition system, different neural network architectures can be employed, e.g. a recurrent neural network or a multi-layer perceptron (MLP). The nets usually

take an acoustic feature vector plus additional context from a number of surrounding frames as input, and output phoneme posterior probability estimates. In the following decoding stage, HMMs are used to combine frame-based probabilities to carry out word and utterance level decoding. In the case of an MLP, the neural network is trained using the error-back-propagation algorithm (Bourlard and Morgan 1993).

Lexicon. The lexicon (or dictionary as it is often referred to) typically consists of the orthography of words that occur in the training material and their corresponding phonetic transcriptions. During recognition, the phonetic transcriptions in the lexicon function as a constraint which defines the sequences of phonemes that are permitted to occur. The transcriptions can be obtained either manually or through grapheme-to-phoneme conversion.

In pronunciation variation research one is usually confronted with two types of lexica: a canonical (or baseline) lexicon and a multiple pronunciation lexicon. A canonical lexicon contains the normative or standard transcriptions for the words; this is a single transcription per word. A multiple pronunciation lexicon contains more than one variant per word, for some or all of the words in the lexicon.

Language Model. Typical recognizers use n-gram language models. An n-gram contains the prior probability of the occurrence of a word (unigram), or of a sequence of words (bigram, trigram etc.):

$$\text{unigram probability } P(w_i) \tag{1.4}$$

$$\text{and bigram probability } P(w_i|w_{i-1}) \tag{1.5}$$

The prior probabilities (priors) in a language model are often estimated from large amounts of training texts for which there is no corresponding acoustic material, i.e., the training texts consist of text material *only*. In the studies presented in this thesis, this is not the case, as the training material used to train the acoustic models is also employed to estimate the probabilities in the language model (see Section 1.5 for more information on this speech material). This makes it possible to incorporate pronunciation variation in the language models, by estimating prior probabilities for the variants in the training corpus, rather than for the words.

Search algorithm. The search algorithm is used to find the most likely sequence of words through the search space in order to maximize the likelihood of W, given the speech signal (or the corresponding acoustic feature vector sequence).

Within the search strategy, a single-pass or multi-pass search can be employed. In the work presented in this thesis, only single-pass search strategies have been employed. However, it has been shown that multi-pass searches can be very useful for pronunciation modeling, as this makes it possible to dynamically change the lexicon. Factors such as rate of speech or type of dialect, which are measured or estimated in a first pass, can be used to determine the appropriate set of pronunciations to include in the

lexicon. This dynamically adjusted lexicon can then be employed in a second pass. Examples of pronunciation variation research in which a multi-pass approach has been used are Fosler-Lussier (1999) and Lee and Wellekens (2001).

Recognized words. The output of the search algorithm is an ordered n-best list of possible hypotheses for the utterance under investigation. The top of the list is compared to a reference transcription to determine the word error rate (WER). The WER is defined as:

$$\text{WER} = \frac{I + D + S}{N} \times 100 \tag{1.6}$$

where I is the number of insertions, D the number of deletions, S the number of substitutions, and N is the total number of words.

Recognized phones. A variant of word recognition is phone recognition. In this type of recognition task, the lexicon does not contain words, but instead contains a list of phones, and a *phone* bigram language model is used to provide phonotactic constraints. The output is a sequence of phones, and instead of a WER, a phone error rate (PER) can be calculated to measure the performance of the system.

Forced recognition. A special type of "recognition" is often employed to automatically obtain transcriptions of the pronunciation variants in the training material, i.e. forced recognition, also referred to as forced Viterbi alignment or forced alignment. In a forced alignment, the recognizer is provided with the orthographic transcription of the material which is to be recognized. Viterbi decoding is used to find the most likely string of phones that match the supplied words, given the acoustic input and various transcriptions for each word. This leads to a new set of time-aligned phonetic labels for the material. Subsequently, these new transcriptions can be used for acoustic model training, and they can also be employed to estimate priors for the language models. Forced alignment is also used as a tool for obtaining information about which pronunciation variation is present in the data (in Section 1.4 this is described in more detail).

Phicos and ICSI recognition systems

Two continuous speech recognition (CSR) systems for Dutch are used in the publications in this thesis: the Phicos recognition system (Steinbiss et al. 1993), and the ICSI hybrid ANN/HMM speech recognition system (Bourlard and Morgan 1993). The main differences between the Phicos and ICSI systems are the search strategies that are used and the manner in which the acoustic probabilities are estimated. The ICSI system uses stack decoding and neural networks are employed to estimate the acoustic probabilities, whereas in the Phicos system, a Viterbi beam search is employed and mixtures of Gaussians are used to estimate the acoustic probabilities.

In both systems, 37 phones were employed to describe the Dutch VIOS data.[1] For the allophones of /l/ and /r/ a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). The other 33 phonemes were context-independent. Models for non-speech sounds and silence were also incorporated in the two ASR systems. Appendix A gives an overview of the phone symbols that were used.

The systems use word-based unigram and bigram language models. The lexicon is the same in both systems, in the sense that it contains the orthography of the words and phone transcriptions for the pronunciations. However, it differs in the sense that the ICSI lexicon also contains prior probabilities for the variants of the words, whereas the Phicos lexicon does not. In the ICSI lexicon the prior probabilities are distributed over all variants for a word and add up to 1.0 for each word. Depending on the type of variants (knowledge-based or data-derived[2]) the prior probabilities are distributed either equally over the variants of a word or they differ for the variants of a word as they are estimated from the training data.

In the Phicos recognition system (Steinbiss et al. 1993), continuous density hidden Markov models (HMMs) with 32 Gaussians per state are used. The HMMs have a tripartite structure, and each of the three parts consists of two states with identical emission distributions. The transition probabilities, which allow for loops, jumps and skips, are tied over all states. Feature extraction is carried out every 10 ms for 16 ms frames. The first step in the feature analysis is an FFT analysis to calculate the spectrum. In the following step, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied to the log of the filterband coefficients. The final processing stage is a running cepstral mean subtraction. In addition to the 14 cepstral coefficients, 14 delta coefficients are calculated, which makes a total of 28 feature coefficients, which are used to describe the speech signal.

The neural network in the ICSI hybrid HMM/ANN speech recognition system (Bourlard and Morgan 1993) was bootstrapped using segmentations of the training material obtained with the Phicos system. These segmentations were obtained by performing a Viterbi alignment using a baseline lexicon (only canonical pronunciations) and Phicos baseline acoustic models, i.e. no pronunciation variation had been explicitly modeled. The front-end acoustic processing consisted of calculating 12^{th}-order PLP features (Hermansky 1990), and energy every 10 ms, for 25 ms frames . The neural net takes an acoustic feature vector plus additional context from eight surrounding frames of features at the input, and outputs phone posterior probability estimates. The neural network has a hidden layer size of 1000 units and the same network was employed in all experiments.

1.3 Pronunciation modeling for ASR

The objective of ASR is to derive the correct string of spoken words from an acoustic signal. However, pronunciation variation makes it more difficult to achieve this objective, as the

[1] See Section 1.5 for a description of the VIOS speech material.

[2] In Section 1.4.1, knowledge-based and data-derived approaches to generating variants are discussed in more detail.

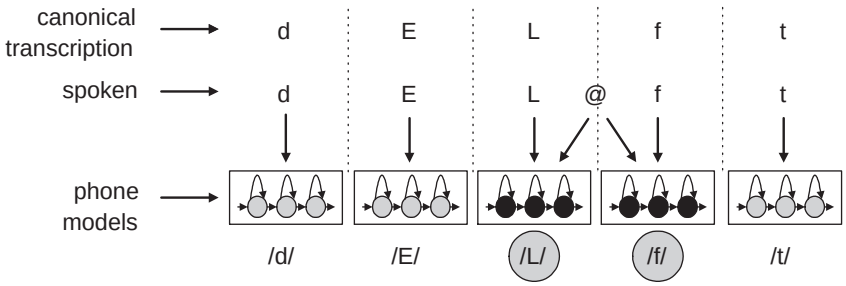


Figure 1.2: Contamination of phone models caused by a mismatch between the acoustic signal and the corresponding transcription during training due to schwa-insertion.

variation can result in recognition errors. The goal of pronunciation modeling is to solve the recognition errors due to pronunciation variation and thus to improve the performance of the ASR system. This section illustrates in what way pronunciation variation can be detrimental to speech recognition both during the training phase and during recognition.

In ASR, the continuous speech signal is described as a sequence of discrete units, which in general are phones.[3] In the studies presented in this thesis, we deal with pronunciation variation that becomes apparent in a careful phonetic transcription of speech, in the form of insertions, deletions or substitutions of phonemes relative to the canonical transcription of the words. This type of pronunciation variation can be said to occur at the segmental level. All of the variation that takes place below the level of the phonetic transcription (for example, the variation due to vocal tract differences) is implicitly left to the HMMs or the neural nets to model.

Figures 1.2 and 1.3 exemplify the way in which pronunciation variation at the segmental level causes problems for ASR during **training**, and consequently why it should be modeled. These figures show how phone models become contaminated when a word's pronunciation differs from the canonically *expected* pronunciation. The first example illustrates the effect of an insertion, and the second example illustrates a deletion. The resulting phone models are contaminated due to the mismatch between the acoustic signal and the phoneme label assigned to it, indicated by the darker grey color of the phone models.

In the example in Figure 1.2, the word "Delft" (Dutch city name) with its canonical transcription /dELft/ [4] is pronounced as /dEL@ft/, i.e., schwa-insertion has taken place. This means that, during training, parts of the acoustic signal corresponding to /@/ are used to train models for /L/ and /f/, causing contamination of the models for /L/ and /f/. In the example in Figure 1.3, "latere" (later) with its canonical transcription /la:t@r@/ is pronounced as /la:tr@/, i.e., schwa-deletion has taken place. During training this leads to contamination of the /@/ model.

[3] For a discussion of the drawbacks of using phonetic transcriptions, see Section 1.7.

[4] SAMPA-notation is used throughout this thesis: `http://www.phon.ucl.ac.uk/home/sampa/dutch.htm`.

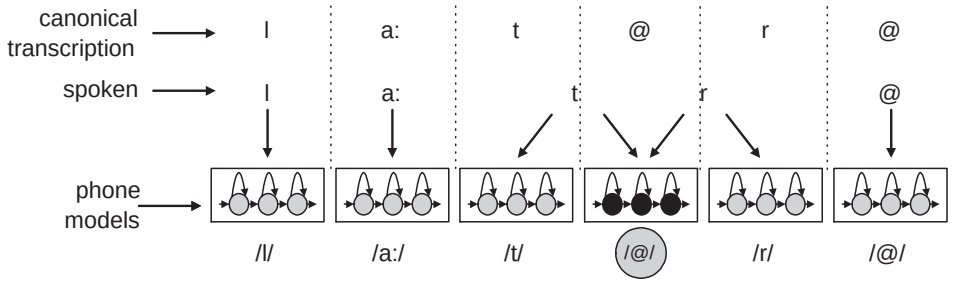


Figure 1.3: Contamination of phone models caused by a mismatch between the acoustic signal and the corresponding transcription during training due to schwa-deletion.

It should be noted that Figures 1.2 and 1.3 are somewhat simplified illustrations of what occurs in reality. In reality, the contamination will not be confined to the phones directly bordering on the deleted or inserted phone, but phones farther removed from the deleted or inserted phone may also be influenced and possibly contaminated. However, contamination of the acoustic models is not *intrinsically* detrimental to speech recognition. Speech, in actuality, is not a sequence of phones strung together like beads-on-a-string with clear-cut boundaries between the individual phones (as Figures 1.2 and 1.3 may falsely suggest). Phenomena such as coarticulation, transitions between phones, feature spreading and cue trading all play a role at or over phone boundaries. Inherently, these phenomena are responsible for a certain degree of contamination of the phones. This type of contamination, however, in contrast to the contamination illustrated in the figures, enhances the capability of the system to cope with "reality". This line of thought also holds for many of the substitutions that take place in speech, as we assume that their properties are also captured implicitly in the acoustic models.

During **recognition**, pronunciation variation may also cause errors. The recognition errors can be a direct result of the fact that contaminated models are less effective in distinguishing between different phones. Another reason why errors may occur is that variants can be pronounced which are not included in the lexicon. For instance, if /la:tr@/ is pronounced but the baseline transcription is /la:t@r@/, the possibility exists that the baseline transcription of another word in the lexicon will match the acoustic signal better, for example, /la:tst@/ ("laatste" meaning "last").

Taking all of this into account, one may wonder whether modeling pronunciation variation at a segmental level can contribute to the improvement of recognition performance. Studies by McAllaster et al. (1998) and Saraçlar et al. (2000) have shown that large improvements are feasible, if there is a match between the acoustic models used during recognition and the transcriptions in the lexicon. In other words, these experiments show that substantial improvements through pronunciation modeling are possible *in principle*.

In McAllaster et al. (1998) simulation experiments were carried out to determine the effect on recognition performance if all of the pronunciation variants encountered by the

decoder were in fact contained in the lexicon. The simulation experiments show that when the data complies perfectly with the probability assumptions of the model (achieved by fabricating the data on the basis of the models) the WER drops from ca. 40% to less than 5%.

In Saraçlar et al. (2000) cheating experiments were conducted by carrying out an unconstrained phone recognition on the *test* speech. The phone string that resulted from this phone recognition was aligned with the reference word transcriptions for the test set and the *observed* pronunciation of each word in the test set was extracted. Next, the pronunciation dictionary was modified individually for each test utterance by including only the *observed* pronunciations for each of the words in the utterance. Using the modified lexicon to rescore a lattice obtained with the baseline ASR system led to a relative improvement of 43% in WER. Both these studies show that the performance can improve substantially if there is a close match between the acoustic models and the transcriptions, in other words, knowing the correct pronunciations can result in large gains.

In a nutshell, the reason for carrying out pronunciation modeling is because of the mismatch between the acoustic signal and the transcription of the signal (i.e. phone transcriptions in the lexicon). During training this mismatch leads to contamination of the acoustic models. Although part of the contamination may be advantageous to speech recognition, there is also a part which may be detrimental to ASR. Neutralizing the problematic contamination in the acoustic models is one of the goals of the work presented in this thesis. In addition, an attempt is made to reduce the mismatch by ensuring that the different pronunciations of a word are accounted for during recognition. In the following section, attention is paid to the methods that are employed to minimize the mismatch between acoustic models and transcriptions, for instance by including the "correct" variants in the lexicon and by removing the contamination from the acoustic models.

1.4 Issues in pronunciation variation modeling

This section gives a description of the issues that play a role when performing pronunciation variation modeling for ASR. It is intended as an introduction to the main approaches in pronunciation variation modeling, to set the scene for the summaries of the publications. For a more comprehensive overview of the approaches to modeling pronunciation variation (and all major references), see Strik and Cucchiarini (1999).

There are two questions which cover most of the issues that must be addressed when modeling pronunciation variation:

1. How is the information obtained that is required to describe pronunciation variation?

2. How is this information incorporated in the ASR system?

In the following two sections these questions are addressed. In Section 1.4.1, the approaches to obtaining information are discussed, and in Section 1.4.2 how it is incorporated.

1.4.1 Obtaining information

Information about pronunciation variation can be acquired from the data itself or through (prior) knowledge; also termed the data-derived and the knowledge-based approaches to modeling pronunciation variation. One can classify approaches in which information is derived from phonological or phonetic knowledge and/or linguistic literature (Cohen 1989; Giachin et al. 1991) under knowledge-based approaches. Existing dictionaries also fit into this category (Lamel and Adda 1996; Roach and Arnfield 1998). In contrast, data-derived approaches include methods in which manual transcriptions of the training data are employed to obtain information (Riley et al. 1999; Saraçlar et al. 2000), or automatic transcriptions are used as the starting point for generating lists of variants (Fosler-Lussier 1999; Wester and Fosler-Lussier 2000).

Although the above approaches are useful, to a certain extent, for generating variants, they all have their drawbacks too. The linguistic literature, including pronunciation dictionaries are not exhaustive; not all processes that occur in spontaneous speech (or even read speech) are described in the linguistic literature, or are present in pronunciation dictionaries. Furthermore, a knowledge-based approach runs the risk of suffering from discrepancies between theoretical pronunciations and phonetic reality. A drawback of hand-transcribed data is that it is labour intensive, and therefore expensive. As a consequence, in general there is rarely sufficient hand-transcribed data. Moreover, manual transcriptions tend to contain an element of subjectivity. Transcriptions made by different transcribers, and even made by the same transcriber, may differ quite considerably (Shriberg and Lof 1991; Cucchiarini 1993). The main problem that is introduced with automatic methods is that phone recognition is not completely reliable either, i.e., it contains errors. This can lead to the generation of pronunciation variants that are the result of mistakes in the recognition, instead of being based on real pronunciation variation.

The options for incorporating the information into the ASR system are determined by the manner in which the variants are obtained. Using theoretical phonological rules limits the possibilities one has to merely adding variants, whereas a manual or good quality automatic transcription allows for more options. In the studies presented in this thesis both major approaches to obtaining variants have been used. In Kessens, Wester, and Strik (1999a) (publication 1), a knowledge based approach to obtaining pronunciation variants for Dutch is investigated. In Wester (2001) (publication 3), in addition to the knowledge-based approach, a data-derived approach is studied. In this study, a comparison is also made between the two approaches by analyzing the degree of overlap between the different lexica they produce.

1.4.2 Incorporating the information in ASR

After the pronunciation variants are obtained, the next question that must be addressed is how the information should be incorporated into the ASR system. There are different levels at which this problem can be addressed. In Strik and Cucchiarini (1999) a distinction was made among incorporating information on pronunciation variation in the lexicon, the acoustic models and the language models. In the following sections, pronunciation modeling at each

of these levels is discussed. First, adding variants to the lexicon is addressed. This is followed by a discussion of lexical confusability, which is an issue that is closely linked to modeling pronunciation variation in the lexicon. Next, the role of forced alignment in pronunciation modeling is explained, before discussing how pronunciation variation can be incorporated in the acoustic models and how the language models are employed in pronunciation modeling. The final issue that is addressed in this section is the use of articulatory-acoustic features in pronunciation modeling.

Adding variants to the lexicon

As speech recognizers make use of a lexicon, pronunciation variation is often modeled at the level of the lexicon. Variation that occurs within a word can be dealt with in the lexicon by adding variants of the words to the lexicon. Variants of a single word are different phonetic transcriptions of one and the same word; i.e., substitutions, insertions and deletions of phones in relation to the base-form variant. This type of variation is within-word variation. However, in continuous speech a lot of variation occurs over word boundaries. This is referred to as cross-word variation. Cross-word variation can, to a certain extent, be dealt with in the lexicon by adding sequences of words which are treated as one entity, i.e., multi-words. The variation in pronunciation that occurs due to cross-word variation is modeled by adding variants of the multi-words to the lexicon (Sloboda and Waibel 1996; Fosler-Lussier and Williams 1999). An alternative method for modeling cross-word variation in the lexicon is described in Cremelie and Martens (1999): the cross-word variants are coded in the lexicon in such a way that during recognition only compatible variants can be interconnected. The importance of cross-word variation modeling was illustrated in Yang and Martens (2000) (the follow-up study to Cremelie and Martens (1999)) which shows that almost all the gain (relative improvement of 45% in WER over baseline performance) in their method is due to modeling cross-word variation.

In most approaches, the lexicon is static, in the sense that it is not altered during the recognition phase. However, there have also been a few studies in which the lexicon was dynamically altered. For instance, Fosler-Lussier (1999) showed that improvements can be found by a dynamic rescoring of n-best lists using a word-based decision tree dictionary. In Lee and Wellekens (2001), a two-pass approach to modeling pronunciation variation is used in which the recognition lexicon is dynamically adjusted depending on the utterance which is being recognized. For further details on pronunciation modeling at the lexical level see Strik (2001).

Lexical confusability

Variants are added to the lexicon to increase the chance that one of the transcriptions of a word will match the corresponding acoustic signal. However, the other side of the coin is that adding variants increases lexical confusability. It has been shown in many studies that simply adding variants to the lexicon does not lead to improvements, and in many cases even causes deteriorations in WER. For instance, in the studies of Yang and Martens

(2000) and Kessens et al. (2001) it was shown that when the average number of variants per word in the lexicon exceeds roughly 2.5, the system with variants starts performing worse than the baseline system without variants. Predicting which pronunciations will be the correct ones for recognition goes hand in hand with dealing with lexical confusability. The dynamic lexica described in the previous section were developed with exactly this problem in mind: dynamically adjusting the lexicon for the utterance that is being recognized should circumvent most of the lexical confusability that is otherwise introduced.

Confusability in data-derived approaches is often introduced by errors in phonetic transcriptions. These phonetic transcriptions are used as the information source from which new variants are derived. Consequently, incorrect variants may be created. One commonly used procedure to alleviate this problem is to smooth the phonetic transcriptions by using decision trees (D-trees) to limit the observed pronunciation variation (Riley and Ljolje 1996; Fosler-Lussier 1999; Riley et al. 1999; Saraçlar et al. 2000; Robinson et al. 2001). In a D-tree approach, an alignment between a canonical transcription and an alternative transcription is used as the input to build the D-trees. The context used for decision making can include anything from mere left and right neighboring phone identity to information such as lexical stress, position of a phone within the syllable, or finer-grained feature information. Using the D-trees, finite state grammars (FSGs) are generated for the words in the training material. These FSGs are realigned with the acoustic signal. The resulting phone transcriptions can be used to generate a new lexicon. In this way, mistakes in the transcriptions can be filtered out.

Other approaches combat confusability by rejecting variants that are highly confusable on the basis of phoneme confusability matrices (Sloboda and Waibel 1996; Torre et al. 1996). In Holter and Svendsen (1999) a maximum likelihood criterion was used to decide which variants to include in the lexicon. In Wester and Fosler-Lussier (2000) a confusability metric was introduced which was used to discard highly confusable variants. Amdall et al. (2000) propose log-likelihood-based rule pruning to limit confusability. Measures such as absolute or relative frequency of occurrence have also been employed to select rules or variants (Cremelie and Martens 1999; Kessens et al. 2001). Finally, confidence measures have been employed to combat confusability by augmenting a lexicon with variants using a confidence-based evaluation of potential variants (Williams and Renals 1998; Fosler-Lussier and Williams 1999).

Both within-word and cross-word variation are investigated in Kessens, Wester, and Strik (1999a) (publication 1). In this study, lexical confusability is not addressed as such, but an analysis is carried out in an attempt to find tools which can be used to decide which variants to add to a lexicon and which ones to leave out. In Wester (2001) (publication 3), the D-tree approach is employed to smooth automatically obtained phone transcriptions. In addition, the confusability metric introduced in Wester and Fosler-Lussier (2000) is further examined as a tool for discarding highly confusable variants.

Forced recognition

Forced recognition (cf. Section 1.2) is employed in various ways in pronunciation modeling. The main objective of using forced alignment in pronunciation modeling is to “clean up”

the transcriptions in the training material, i.e., to obtain a more precise transcription given multiple transcriptions for the words in the lexicon. In the data-derived D-tree approach forced alignment is used to align the FSGs with the training data; to subsequently select variants on the basis of the output of the alignment. The alignments are also used to obtain priors for the pronunciation variants in the lexicon, or to estimate the probabilities in the language model. Finally, the transcriptions can also be employed to retrain the acoustic models.

In Wester et al. (2001) (publication 3), an explicit investigation into the performance of forced alignment was carried out. The goal of this study was to ascertain how reliably the CSR system performs compared to human listeners with regard to choosing variants.

Acoustic models

The objective of retraining the acoustic models on the basis of the output of forced alignment is not only to obtain more accurate acoustic models but also to achieve a better match between the multiple pronunciation lexicon and the acoustic models used during recognition. In various studies improvements in recognition results were found after retraining the acoustic models (Sloboda and Waibel 1996; Riley et al. 1999). However, in some studies no difference in performance was measured (Holter and Svendsen 1999), or even a deterioration was found (Beulen et al. 1998). Strik and Cucchiarini (1999) mention that these retranscription-retraining steps can be iterated, and Saraçlar (2000) and Kessens et al. (1999b) demonstrate that most of the gain is found as a result of the first iteration.

Optimizing the acoustic models so that they better match the transcriptions is one way to reduce the mismatch between the acoustic models and the transcriptions. Other approaches have also been taken in which the lexicon is left unchanged and the pronunciation deviations are reflected in the acoustic model topology (Eide 1999; Saraçlar 2000). Examples of methods that explicitly account for coarticulation and transitions between neighboring phones at the acoustic level are the speech production model (Blackburn and Young 1995) or the hidden dynamic model (Richards and Bridle 1999; Picone et al. 1999).

Retraining the phone models is an integral part of the knowledge-based approach to modeling pronunciation variation as implemented in Kessens, Wester, and Strik (1999a) (publication 1). In Wester (2001) (publication 3), the effect of retraining the acoustic models is also investigated.

Variant probabilities

Incorporating pronunciation variation in the language model can be carried out by estimating the probabilities of the variants instead of the probabilities of the words. This is of course only possible if the pronunciation variants are transcribed in the training material, and the language models are trained on this material. An intermediate level of modeling pronunciation variation in the language model is possible in the form of word classes. In particular, this approach is taken to deal with processes of cross-word variation such as liaisons in French (Brieussel-Pousse and Perennou 1999).

Many studies (Cohen 1989; Yang and Martens 2000; Ma et al. 1998; Fosler-Lussier 1999) have shown that probabilities of the variants (or probabilities of rules) play an important role in whether an approach to modeling pronunciation variation is successful or not. Prior probabilities of the variants can be incorporated in the language model or in the lexicon, depending on the type of recognizer that is being used.

Incorporating variants in the language model is an integral part of the method for modeling pronunciation variation reported in (Kessens, Wester, and Strik 1999a) (publication 1). This approach is necessary as in the Phicos recognition system incorporating priors for the variants in the system is only possible through the language model. Incorporating priors for variants in the ICSI system is possible in the lexicon, thus obviating the need for priors of variants in the language model. Experiments investigating the effect of including or excluding priors during recognition are reported in Wester (2001) (publication 3).

Articulatory-acoustic features

Articulatory-acoustic (phonetic) features have been proposed as an alternative means of classifying speech segments (Kirchhoff 1999; Chang et al. 2000; King and Taylor 2000). One of the reasons for using articulatory-acoustic features is that under many circumstances the segmental approach (based on phone sequences) does not incorporate enough detail with which the subtlety and richness in the speech signal can be captured at the phonetic level (Chang et al. 2001). A similar, but distinctly different approach is to employ articulatory features either inferred from the data using linguistic rules (Deng and Sun 1994) or directly employing articulatory datasets (King et al. 2000). A more complete overview of the approaches taken to employing articulatory features in ASR is given in Wrench (2000).

An oft mentioned advantage of articulatory-acoustic (phonological) features in speech recognition is that these features are better suited for pronunciation modeling than a purely phone-based approach. Few studies, however, have investigated whether this claim is justified or not. In a recent study (Lee and Wellekens 2001) an approach to modeling pronunciation variation was described in which articulatory-acoustic features are used. Lee's approach consists of generating a multiple variant static lexicon during training, which is dynamically adjusted during recognition. The information used to generate pronunciation variants is obtained by extracting features from the speech signal (using an approach similar to King and Taylor (2000)). The features are mapped to phones which are then connected to each other to build a pronunciation network. All possible pronunciations are generated from the network and the output is smoothed by a two-pass forced recognition. The remaining variants are stored in the static lexicon. During recognition this static lexicon is adjusted per utterance. Articulatory-acoustic features are extracted from the test material, mapped to phones, and used to select those entries from the static lexicon that best match the phonetic characteristics of a given speech signal. The selected entries constitute the dynamic lexicon, which is used for recognition. A 16% relative reduction in WER was found on TIMIT (Lamel et al. 1986) compared to their baseline system.

Another advantage of articulatory-acoustic features, which is often mentioned in academic literature, is that the models based on them should generalize better across languages.

In Wester, Greenberg, and Chang (2001) (publication 4), experiments are described to analyze how well features trained on English data perform on Dutch data, in order to ascertain to what extent cross-linguistic transferability of features is feasible.

1.5 Speech material

Before summarizing the articles that form the body of this thesis, a short description is given of the speech material which was used in the studies. The speech material was gathered by recording calls to the on-line version of a spoken dialogue system entitled OVIS (Strik et al. 1997). Thus, the speech consists of extemporaneous, prompted human-machine telephone dialogues. OVIS is employed to automate part of an existing Dutch public transport information service. A large number of telephone calls of the on-line version of OVIS have been recorded and are stored in a database called VIOS. The speech material consists of interactions between human and machine and the data clearly show that the manner in which people speak to OVIS varies, ranging from hypo-articulated speech to hyper-articulated speech.

VIOS speech material is used in all of the studies included in this thesis. The material (3531 dialogues) was divided into a portion for training which consists of 25,104 utterances (81,090 words) and a portion for testing which consists of 6,267 utterances (20,489 words). This corresponds to a total duration of 24h, of which 10.8h is speech and 13.2h is silence. Approximately 60% of the speakers are male and 40% are female. Recordings with a high level of background noise were excluded.

Figure 1.4 shows the cumulative frequency of occurrence of the words in the VIOS training material as a function of word frequency rank. This figure gives an impression of the composition of the VIOS material. Figure 1.4 shows that roughly 80% of the training material is covered by the 100 most frequently occurring words. In total, 1104 unique words occur in the training material. The 14 most frequently observed words are all one syllable long and cover 48% of the training material. Furthermore, as the VIOS corpus comprises data collected from a train timetable information system, 43% of the words in the lexicon concern station names, which corresponds to 16% of the words in the training material.

The transcriptions for the baseline lexicon, which contains *one* variant per word, were obtained using the transcription module of a Dutch Text-to-Speech system (Kerkhoff and Rietveld 1994), which looks up the words in two lexica: CELEX (Baayen 1991) and ONOMASTICA, which was used specifically for station names (Quazza and van den Heuvel 2000). For those words for which no transcription was available, a grapheme-to-phoneme converter (Kerkhoff and Rietveld 1994) was used. All transcriptions were manually checked and corrected when necessary.

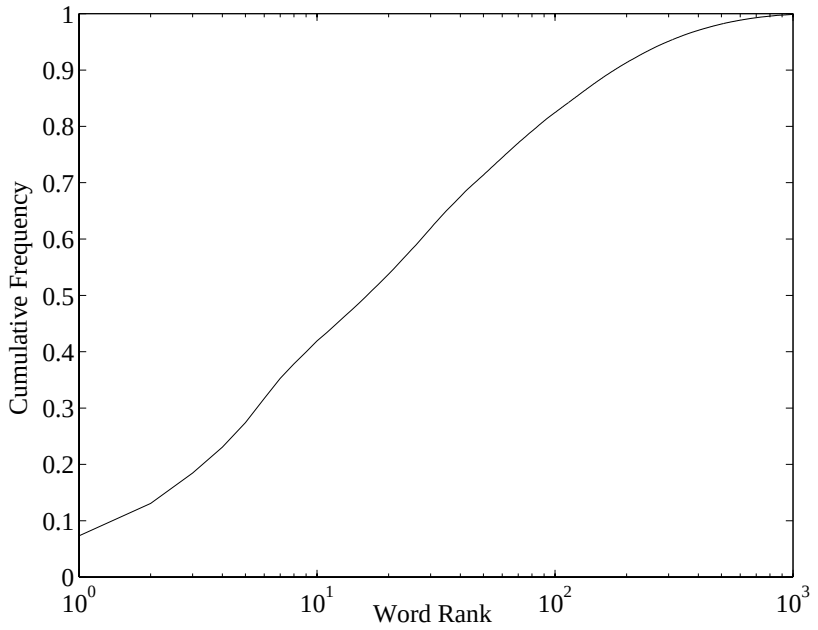


Figure 1.4: Cumulative frequency of occurrence as a function of word frequency rank for the words in the VIOS training material.

1.6 Summary of publications

This section contains the summaries of the four publications contained in Part II of this thesis.

1.6.1 Summary 1: A knowledge-based approach to modeling pronunciation variation for Dutch

J.M. Kessens, M. Wester and H. Strik (1999) Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication, 29*, 193-207.

In this article a description is given of how the performance of a Dutch continuous speech recognizer (CSR) was improved by modeling pronunciation variation using a knowledge-based approach. The objective of the article was to develop a method for modeling Dutch pronunciation variation which could be used to tackle the problem of pronunciation variation for Dutch CSR. Our long term goal was to find the set of rules which is optimal for modeling pronunciation variation. In addition, we were interested to determine whether the trends in

recognition results measured when testing different sets of variants in isolation are the same as those obtained when testing them in combination. In other words, we wanted to answer the question of whether the sum of the effects of sets of variants in isolation is the same, or almost the same, as the total effect of the combination of the sets of variants.

In order to achieve this objective, we proposed a general procedure for modeling pronunciation variation. This procedure affects all three levels of the CSR at which modeling can take place: i.e. the lexicon, the phone models and the language model (Strik and Cucchiarini 1999). This means that variants were added to the lexicon and language models, and that the phone models were retrained on a retranscription of the training material obtained through forced alignment. The general procedure was employed to model within-word variation as well as cross-word variation.

Within-word pronunciation variants were generated by applying a set of five optional phonological rules to the words in the baseline lexicon. The five phonological rules were /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion. These rules were tested in isolation and in combination.

A limited number of cross-word processes were modeled, using two different techniques. The type of cross-word processes we focussed on were cliticization, reduction and contraction (Booij 1995). The first technique consisted of modeling cross-word processes by adding the cross-word variants directly to the lexicon (cross-word method 1), and in the second approach this was done by using multi-words (cross-word method 2). These cross-word approaches were each tested in isolation and in combination with the set of within-word variants (all five rules).

The main results that we found are the following. The baseline system WER is 12.75%. For the within-word method, adding pronunciation variants to the lexicon leads to an improvement of 0.31% compared to the baseline. When, in addition, retrained phone models are used, a further improvement of 0.22% is found, and finally, incorporating variants into the language model leads to a further improvement of 0.15%. In total, a small but statistically significant improvement of 0.68% was found for modeling within-word pronunciation variation.

Each of the phonological rules was also tested in isolation by adding the variants to the lexicon. We found that the rule for /n/-deletion leads to an improvement. The variants generated by the rules for /r/-deletion and /@/-deletion seem to have almost no effect on WER at all. The variants for /t/-deletion and /@/-insertion lead to deteriorations in WER compared to the baseline. The sum of these results is a deterioration in WER of 0.02%, whereas combining the five rules leads to an improvement of 0.31% compared to the baseline.

Using the methods for modeling cross-word pronunciation variation, a total improvement of 0.16% was found for cross-word method 1, and 0.30% for cross-word method 2. A combination of modeling within-word and cross-word pronunciation variation leads to a total improvement of 0.61% for method 1, and a total improvement of 1.12% for cross-word method 2. However, a great deal of the improvement for cross-word method 2 is due to adding multi-words (0.34%). We also investigated whether the sum of the improvements for the cross-word methods tested in isolation is comparable to the improvement obtained

when testing combinations of the methods, and found that this is not the case. For cross-word method 1, the sum of the methods in isolation gives better results than using the methods in combination, whereas for cross-word method 2, the combination leads to larger improvements than the sum of the results in isolation.

On the basis of the results, we concluded that it is clear that the principle of superposition does not apply, neither for the five rules of the within-word method nor for the within-word method in combination with each of the two cross-word methods. The implication of these findings is that it does not suffice to study sets of variants in isolation. Instead, they have to be studied in combination. However, this poses a practical problem as there are many possible combinations.

To further understand the results that were found, we carried out a partial error analysis in which the utterances recognized with the baseline system were compared to those recognized with the experimental condition in which pronunciation variation was incorporated at all levels for a combination of within-word variants and cross-word variants modeled by multi-words. This error analysis showed that 14.7% of the recognized utterances changed, whereas a net improvement of only 1.3% in the sentence error rate was found (and 1.12% in the WER). Thus, the WER only reflects the net result obtained, and our error analysis showed that this is only a fraction of what actually happens due to applying our methods.

To summarize, we obtained the best results when within-word pronunciation variation and cross-word pronunciation variation using multi-words were modeled in combination, and all the steps of the general procedure had been carried out. Using only five phonological rules and 22 multi-words, a relative improvement of 8.8% was found (12.75% - 11.63%).

1.6.2 Summary 2: Forced recognition versus human listeners

M. Wester, J.M. Kessens, C. Cucchiarini and H. Strik (2001) Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. *Language and Speech, 44*(3), 377-403.

The aim of this research was to determine whether the forced recognition technique that we used in our pronunciation variation research could also be used meaningfully, in spite of its limitations, to obtain phonetic transcriptions for linguistic research. In the last decade an increasing number of databases have been recorded for the purpose of speech technology research. These databases contain a wealth of information concerning human language and speech, which makes them very interesting for use in linguistic research. However, before the speech material contained in the databases can be used for phonetic research it has to be phonetically transcribed. The problem is that obtaining good manual phonetic transcriptions is time-consuming, expensive and tedious. Therefore, it would be useful if the transcriptions could be obtained automatically. An automatic way of obtaining a representation that approaches phonetic transcription is using forced recognition (or forced alignment).

In forced recognition, the CSR is constrained by only allowing it to recognize the words present in the utterance being recognized. To this end, the orthographic transcription of

the utterance is needed. The forced choice entails choosing between several pronunciation variants for each of the words present in the utterance, thus leading to a transcription which is more accurate than a simple canonical word-level transcription.

In this study, two experiments were performed in which different comparisons were carried out between the automatically obtained transcriptions and the transcriptions made by human transcribers. The speech material was selected from VIOS. The processes we studied were insertions and deletions of phones. Variants were generated using the same five phonological rules as in Kessens, Wester, and Strik (1999a). Given that there is no absolute truth concerning the question of what phones a person has produced, there is also no reference transcription that can be considered correct and with which the automatic transcription can be compared (Cucchiarini 1993, pp. 11-13). To try and circumvent this problem as much as possible, we used the two most common approaches to obtaining a reference transcription: the majority vote procedure and the consensus transcription.

In the first experiment, four types of comparisons were made to study how the machine's performance relates to that of nine expert listeners. The task, which was exactly the same for the CSR and the listeners, was to decide whether a segment (an /n/, /r/, /t/ or /@/) was present or not in 467 cases.

First, the degree of agreement in machine-listener pairs was compared to the degree of agreement in listener-listener pairs. Degree of agreement is expressed using Cohen's kappa (κ). We found that there is a great deal of variation among the various listener pairs: the listeners' κ values vary between 0.49 and 0.73, and the median for all listener pairs is 0.63. The agreement values for the listener-CSR pairs vary between 0.52 and 0.60, and the median κ value is 0.55. Statistical tests showed that the CSR and three of the listeners behave significantly differently from the other listeners. The agreement for the CSR and one of the listeners is significantly lower than the rest of the listeners, whereas for two other listeners agreement is significantly higher, thus, leaving a middle group of 6 listeners that do not significantly differ from each other.

Second, in order to be able to say more about the quality of the machine's transcriptions and the transcriptions made by the nine listeners, we compared all of the transcriptions to a reference transcription (majority vote procedure). The reference transcription based on the majority vote procedure is stricter when more of the listeners agree. We found that the degree of agreement between the reference transcription and both the CSR and the listeners gradually increases as the reference transcription becomes stricter.

Third, because it can be expected that not all processes give the same results, the comparisons with the reference transcription were carried out for each individual process of deletion and insertion. This comparison showed that there is no significant difference between the listeners and the CSR for /r/-deletion and schwa-insertion. For the other three processes the differences were significant. Apparently, it is not only the sound in question that counts, be it an /n/ or a schwa, but rather the process being investigated. This is borne out by the fact that the results are so different for schwa-deletion as opposed to schwa-insertion.

Fourth, a more detailed comparison of the choices made by the machine and by the listeners was carried out to get a better understanding of the differences between the ma-

chine's performance and that of the listeners. These experiments showed that across-the-board the listeners registered more instances of insertion and fewer instances of deletion than the machine did, thus showing a stronger tendency to perceive the presence of a phone than the machine. Although this finding was consistent over the various processes, it was most pronounced for schwa-deletion.

A second experiment was carried out in order to find out why and in what way the detection of a phone is different for the CSR and for the listeners. In order to study this, a more detailed reference transcription was needed. Therefore, we used a consensus transcription instead of a majority vote procedure to obtain a reference transcription. The results of the second experiment show that the CSR and the listeners have different durational thresholds for detecting a phone. A different mapping between the machine and the listeners' results brought the degree of agreement between the two sets of data closer to each other.

To summarize, we explored the potential that a technique developed for CSR could have for linguistic research. In particular, we investigated whether and to what extent a tool developed for selecting the pronunciation variant that best matches an input signal could be employed to automatically obtain phonetic transcriptions for the purpose of linguistic research. We concluded that the results of our experiments indicate that the automatic tool proposed in this paper can be used effectively to obtain phonetic transcriptions of deletion and insertion processes, although it remains to be seen whether these techniques can be extended to other processes such as substitutions or other deletion/insertion processes. Furthermore, there are significant differences between the CSR and the listeners, but the differences in performance may well be acceptable, depending on what the transcriptions are needed for. Once again it should be kept in mind that the differences that we found between the CSR and the listeners were also in part found between the listeners.

1.6.3 Summary 3: Knowledge-based and data-derived pronunciation modeling

M. Wester (2001) Pronunciation modeling for ASR – knowledge-based and data-derived methods. *Submitted to Computer Speech and Language*.

In this paper, we report on two different approaches to dealing with pronunciation variation: a knowledge-based and a data-derived approach. These approaches differ in the way that information on pronunciation variation is obtained. The knowledge-based approach consists of using phonological rules to generate variants. The data-derived approach consists of performing phone recognition, followed by smoothing using decision trees (D-trees) to alleviate some of the errors in the phone recognition.

The first objective was to compare these two methods of modeling pronunciation variation. In addition to comparing the WER results, the lexica obtained through the different approaches were investigated, to analyze how much of the same pronunciation variation the approaches were modeling.

The second objective was to decide which variants to include in the lexicon and which

ones to exclude. This issue was dealt with by using a confusability metric (introduced in Wester and Fosler-Lussier (2000)) to measure the degree of confusability in a certain lexicon, and also to discard highly confusable variants.

The third objective in this study was to determine whether WER results obtained with a certain lexicon are possibly recognizer dependent. Especially in a data-derived approach, the question arises as to whether pronunciation variation is truly being modeled, or if the system is merely being tuned to its own idiosyncrasies.

The two recognition systems we used are the ICSI hybrid HMM/ANN speech recognition system (Bourlard and Morgan 1993) and the Phicos recognition system (Steinbiss et al. 1993). The baseline results of the two systems on the VIOS material were similar and significantly better than the baseline result that was reported for the Phicos system as employed in Kessens, Wester, and Strik (1999a). The improvement is due to using 12^{th}-order PLP features instead of 14 MFCCs, and employing extra context information.

The knowledge-based approach in this study was very similar to the approach described in Kessens, Wester, and Strik (1999a) although no cross-word pronunciation modeling was carried out. To recapitulate, five optional phonological rules were applied to the words in the baseline lexicon (/n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion), and all the variants generated by the rules were added to the baseline lexicon.

The data-derived approach we used is based on the decision-tree (D-tree) pronunciation modeling approach developed by Riley and Ljolje (1996). In this approach, first of all, phone recognition is carried out on the training material to supply the raw information on pronunciations. Next, an alignment between the phone recognition output and a reference (canonical) transcription is made. A distance measure based on binary phonetic features is used to align the strings of phones and to insert word boundaries at the most appropriate places in the string. At this point, a lexicon is created by adding all the variants to the lexicon; this lexicon is referred to as the phone recognition lexicon. In the D-tree approach, D-trees are used to smooth the phone recognition output before generating a lexicon. We use relatively simple D-trees, only taking into account the identity of the left and right neighboring phones, and the position of the phone within the syllable. For each of the 37 phones (and for the noise model) a D-tree was built. The D-tree model is trying to predict:

$$P(realization \mid canonical, context) \tag{1.7}$$

by asking questions about the context. Using the distributions in the D-trees, finite state grammars (FSGs) were built for the utterances in the training data. During this FSG construction, transitions with a probability lower than 0.1 were disallowed. Subsequently, the FSGs were realigned with the training data, and the resulting phone transcriptions were used to generate a new lexicon.

The confusability of individual variants in a lexicon and the overall confusability in a lexicon were determined on the basis of a forced alignment of the training data using the lexicon for which confusability was to be determined. The resulting phone transcription of the training material is matched to all the words in the lexicon, producing a lattice of words which contains the set of variants that matches any substring within the phone transcription.

On the basis of this lattice, the overall confusability in the lexicon is calculated by adding up the number of variants that correspond to each phone, divided by the total number of phones in the training material. Word level confusability scores are obtained by counting the number of times a variant of a certain word matches the phone transcription of other words in the training material. Those variants which were earmarked by the confusability metric as highly confusable were discarded from the lexicon.

Our first objective was to compare knowledge-based and data-derived approaches to modeling pronunciation variation. Using the ICSI system to carry out the experiments, we found no improvement over the baseline result when the five phonological rules were used to model pronunciation variation. Adding all the variants from the raw phone recognition to the baseline lexicon led to a deterioration in performance. Modeling pronunciation variation using D-trees led to a statistically significant improvement in the ICSI system. A relative improvement of 7.5% compared to the baseline result was found.

Employing the Phicos system to carry out the experiments led to roughly the same degree of improvement for both approaches (3% for the knowledge-based approach and 4% for the data-derived approach). The improvement for the knowledge-based approach was smaller than expected, as in previous work (Kessens, Wester, and Strik 1999a) the improvement due to modeling pronunciation variation had been significant compared to the baseline (relative improvement of 5%). This can be explained by the fact that the starting value of WER in this work is significantly lower than in Kessens, Wester, and Strik (1999a). Our results show that even though the trends are the same, pronunciation modeling through phonological rules has less effect when the WER value is lower to start with. In this case, it seems that part of the mistakes that were previously solved by modeling pronunciation variation are now being taken care of by improved acoustic modeling.

The knowledge-based and data-derived approaches were also compared to each other by analyzing how much overlap exists between the different lexica. Analysis of the lexica showed that the D-trees are, in effect, learning phonological rules. We found that 10% of variants generated by the phonological rules were also found using phone recognition, and this increased to 28% when the phone recognition output was smoothed by using D-trees. Apparently, phonological rule variants are created which were not present in the output of the raw phone recognition. This is a clear advantage of using D-trees over simply using phone recognition output, because the D-trees are capable of generalizing beyond what has been seen in the training material, whereas when the phone recognition approach is employed directly, unseen pronunciations cannot be predicted. Furthermore, it is an indication that pronunciation variation is indeed being modeled.

Confusability is intuitively an extremely important point to address in pronunciation modeling. The confusability metric proved to be useful as a method for pruning variants from the lexicon. The results show that simply pruning highly confusable variants from the phone recognition lexicon leads to an improvement compared to the baseline. In other words, the confusability metric is a very simple and easy way of obtaining a result which is comparable to the result obtained using methods such as phonological rules or D-trees.

We also intended to use the confusability metric to assign a score to a lexicon which could

then be used to predict how well a lexicon would perform. Overall lexical confusability scores showed that the highest degree of confusability is found in the phone recognition lexica; this is followed by the D-trees lexica, and the least amount of confusability is contained in the phonological rule lexica. However, there is no straightforward relationship between the confusability score and the WER performance. Consequently, it is not clear how the confusability score could be used to predict which lexicon is "better". In addition, there is no relationship between the number of entries in the lexicon (or the number of variants per word) and the WER.

One of the questions we were interested in answering was: "Is pronunciation variation indeed being modeled, or are idiosyncrasies of the system simply being modeled?" We found that simply employing the D-trees lexicon (generated using the ICSI system) in the Phicos system led to a significant deterioration in WER compared to the baseline result. For the ICSI system a comparable deterioration was found when the variant probabilities were not taken into account during recognition. When these probabilities were incorporated in the systems the WER improved dramatically in both cases. The similarity in the results obtained using two quite different recognition systems indicate that pronunciation variation is indeed being modeled.

To conclude, a knowledge-based approach for modeling pronunciation variation in Dutch using five phonological rules leads to small improvements in recognition performance. Using a data-derived approach leads to larger improvements when the phone recognition output is either smoothed by D-trees or pruned using the confusability metric. Both techniques result in roughly the same improvement. Furthermore, it is encouraging that using the same pronunciation variant lexicon in two different recognition systems leads to roughly the same results, as this indicates that pronunciation variation is indeed being modeled and not merely the idiosyncrasies of a certain recognition system.

1.6.4 Summary 4: Turning to articulatory-acoustic features

M. Wester, S. Greenberg and S. Chang (2001) A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, pp. 1729-1732.

Current generation ASR systems often rely on automatic alignment of the training material with the acoustic signals to train and refine phonetic segment models. However, the alignments may not be as accurate as desirable, compared to hand transcripts. A potential means to improve automatic transcriptions is through the use of articulatory-acoustic features (AF) instead of phones for classification.

Ultimately, the goal is to achieve improved automatic speech recognition. In this research, we wanted to ascertain whether articulatory-acoustic features trained on English (NTIMIT) data could transfer to Dutch (VIOS) data. We also explored the potential of applying an "elitist" approach for AF classification to Dutch. An advantage of the "elitist" approach is that it provides a potential means of automatically transcribing a corpus at the phonetic level

without recourse to a word-level transcript.

Two separate corpora, one for Dutch, the other for American English, were used in this study. One hour of Dutch VIOS material was selected for training the nets for the classification of articulatory features. The American-English NTIMIT material consisted of roughly three hours of training material. An eighteen-minute component of VIOS that was hand transcribed by students at the University of Nijmegen was used as a test set.

Multi-layer perceptrons (MLP) were trained on five separate feature dimensions: (1) place and (2) manner of articulation, (3) voicing, (4) rounding and (5) front-back articulation. Articulatory-acoustic features were automatically derived from phonetic-segment labels. For example the phone /b/ would receive the labels bilabial, stop, +voice, N/A and N/A (N/A meaning not applicable). The features "rounding" and "front-back" only apply to vowels.

The front-end representation of the signal consisted of logarithmically compressed power spectra computed over a window of 25 ms every 10 ms. The spectrum was partitioned into fourteen, 1/4-octave channels between 0.3 and 3.4 kHz. Delta and double-delta features pertaining to the spectral contour over time were also computed. The outputs from the MLP are articulatory-acoustic features.

Classification of articulatory-acoustic features trained and tested on VIOS was more than 80% correct at frame level for all dimensions except for place of articulation. Overall this performance is comparable to that associated with American English (Chang et al. 2000) and German (Kirchhoff 2000) material.

The results for cross-linguistic classification showed that the classification of a system trained on NTIMIT and tested on VIOS is lower than a system trained and tested on VIOS. The decline in performance is ca. 10-15% (absolute) for all feature dimensions, except for place, for which there is a larger decline. Voicing is the one feature dimension in which classification is nearly as good for a system trained on English as it is for a system trained on Dutch. The manner dimension also transfers reasonably well from training on NTIMIT to VIOS.

Frames situated in the center of a phonetic segment tend to be classified more accurately than those close to the segmental borders. Furthermore, the confidence with which these center frames are classified is higher, especially for the manner of articulation dimension. Therefore, we investigated to what extent classification could benefit from frame selection. By using a network-output threshold of 0.7 for frame selection it is possible to improve the accuracy of manner classification between 5 and 10% (absolute) when training and testing on VIOS. In the cross-linguistic case, training on NTIMIT and testing on VIOS, an improvement in accuracy is found between 1 and 9% (absolute) for the various categories. The overall accuracy at the frame level increases from 73% to 81%. For the stop and nasal categories, the performance does not improve appreciably.

Place of articulation information is of crucial importance for classifying phonetic segments correctly (Greenberg and Chang 2000) and (Kirchhoff 1999). Unfortunately, place of articulation is the most poorly classified of the five feature dimensions. The reason place of articulation is so poorly classified could be the heterogeneous nature of the articulatory-acoustic features involved. Place of articulation for vowels is of a different type altogether

compared to place of articulation for consonants. Moreover, even among consonants, there is a lack of concordance in place of articulation.

Articulatory-acoustic features provide a potentially efficient means of developing cross-linguistic speech recognition systems. The present study demonstrates that certain AF dimensions such as voicing and manner of articulation transfer relatively well from English to Dutch. However, a critical dimension, place of articulation, transfers poorly. An appreciable enhancement of place-of-articulation classification can result from manner-specific training.

1.7 A Critical Appraisal

Since the early 1970s, modeling pronunciation variation in automatic speech recognition has been a topic of interest to researchers in the field of ASR, and a large amount of time and effort has been invested in dealing with the problem of pronunciation variation. However, the improvements in WERs as a result of explicit modeling of segmental variation have not been quite as large as had been expected, as the following citations illustrate:

> "The disparity between improved performance of decision tree classifiers and the lack of large improvements when these models are employed in dynamic rescoring of n-best lists is puzzling."
>
> —Fosler-Lussier (1999, pp. 151)

> "While many studies have pointed to pronunciation variability as a key problem, the work on pronunciation modeling in terms of phone-level substitutions, deletions and insertions has so far only yielded small performance gains."
>
> —Shafran and Ostendorf (2000)

> "There have been a variety of attempts to handle this kind of problem ["going to" being realised as "gonna"] within the *beads-on-a-string* framework [...] eg by using decision trees to generate context dependent pronunciations. However, none have been successful."
>
> —Young (2001)

These quotes illustrate the feeling that is present in the pronunciation modeling community and it is a feeling which contrasts sharply with the best-case-scenario studies (McAllaster et al. 1998; Saraçlar et al. 2000) that suggest that improved pronunciation models should bring much lower WERs than are reported in most pronunciation variation research at present.

In the following sections, I will attempt to summarize the underlying reasons why the improvements are not as large as may have been expected. However, first I would like to mention that there are examples of pronunciation variation modeling where large improvements have been found. For instance in the work carried out by Cremelie and Martens (1999) and Yang and Martens (2000) relative improvements of up to 45% were found and

in Bacchiani and Ostendorf (1999) a 20% reduction in error is reported. Although these are impressive results, it should be noted that the results were found for read speech corpora (Resource Management and TIMIT), and it is not self-evident that these results will scale to more spontaneous speech. In spontaneous speech there is more variation in pronunciation than in read speech (Weintraub et al. 1996), therefore it can be conjectured that there is more room for improvement which possibly could be achieved by pronunciation modeling. However, it is not clear that methods developed for read speech will have the same effect on spontaneous speech. To further exemplify this, Bacchiani and Ostendorf (1999) report that preliminary experiments on spontaneous speech demonstrate only small gains, in contrast to the 20% reduction on read speech mentioned earlier, and that further experiments are necessary.

In my view there are a few clear problems linked to modeling pronunciation variation at a segmental level which are responsible for the limited success of the various methods: viz. lexical confusability, phone transcriptions, and the beads-on-a-string paradigm.

1.7.1 Lexical confusability

It is clear that words can be pronounced in many different ways. It is also clear that this constitutes a problem for speech recognition. The most obvious way of dealing with this variation is to add variants to the lexicon. However, describing pronunciation variation by adding variants to the lexicon leads to an increase in lexical confusability. As mentioned in Section 1.4.2, this problem has been signaled by many in the field of pronunciation modeling, and many different solutions have been suggested for dealing with this problem. Although lexical confusability may present difficulties, it should not be forgotten that it is part and parcel of a lexicon. There will always be confusable word pairs and homophones, simply because they exist in speech and in language.

Despite the increase in lexical confusability caused by adding variants to the lexicon this approach does have merit in the sense that some of the variation in the speech material can be captured and modeled (provided that prior probabilities of the variants are taken into account). The results presented in this thesis show that this is the case for the VIOS database. Furthermore, statistically significant improvements have also been found on corpora such as Switchboard and the Broadcast News corpus (Riley et al. 1999; Fosler-Lussier 1999). However, the effect of adding variants is limited, as the improvements in WER are generally not very large on (semi-)spontaneous speech.

The goal of modeling pronunciation variation is to lower WERs. Simulation studies and cheating experiments (McAllaster et al. 1998; Saraçlar et al. 2000) have shown that if one can accurately predict word pronunciations in a certain test utterance the performance should improve substantially. However, substantial improvements through pronunciation modeling have not yet been achieved. The following explanation clarifies what may be the cause of this lack of improvement. In a lexical approach to pronunciation modeling, the prior probabilities for the variants are usually estimated from the training material, and local context effects are not taken into account. In various studies (Fosler-Lussier and Morgan 1999; Jurafsky et al.

2001), it has been shown that the degree and also type of pronunciation variation for a word depends on the local context of that word. Consequently, it may be that prior probabilities for variants just do not suffice. In addition to the prior probabilities, conditional probabilities for pronunciation variants should be incorporated in the recognition system. If the set of variants which is used during recognition can be dynamically adjusted per utterance by using context information then lexical pronunciation variation may lead to lower WER results. And possibly, improvements such as those reported in simulation studies and cheating experiments (McAllaster et al. 1998; Saraçlar et al. 2000) can be mimicked in real conditions.

1.7.2 The dubious nature of phone transcriptions

In almost all approaches to modeling pronunciation variation, automatic transcriptions play a role. The quality of these automatic transcriptions is usually measured by comparing them to human transcriptions. However, manual phonetic transcriptions tend to contain an element of subjectivity. Therefore, there is no absolute truth as to what phones a speaker has produced in an utterance (Cucchiarini 1993).

A number of recent studies once again show that phonetic transcription of conversational speech is quite difficult for human labelers. For instance, inter-labeler agreement for the Switchboard Transcription Project[5] ranged between 72% and 80% on the phonetic segment level (Greenberg 1999). The transcription of German data showed that transcribers reached an agreement of 93.1% to 94.4% for careful speech and between 78.8% and 82.6% for less careful speech (Kipp et al. 1996; Kipp et al. 1997). Results on our data show that agreement between listeners ranges from 75% to 87% for pairs of listeners (Kessens et al. 1998). Furthermore, Saraçlar and Khudanpur (2000) showed that the inherent ambiguity in the identity of phonetic segments in spontaneous speech makes the notion of phonetic transcription, be it manual or automatic, a difficult one.

These examples all indicate the dubiousness of using phonetic transcriptions to describe speech. Moreover, if human transcribers do not even agree how can the CSR be expected to produce the correct transcription of a speech signal in terms of phones. The fact that human transcribers do not totally agree with each other suggests that phones are sub-optimal units for describing speech, and consequently, perhaps phones are also sub-optimal units for automatic speech recognition. However, having said that, it is not clear what the worthy successor(s) of the phone should be.

1.7.3 Beads-on-a-string

In various papers, the following question has been asked: (paraphrased here) “Why is the recognition performance on spontaneous speech so far below human performance?” All the answers point in the same direction: the failure of the assumption that speech can be described as a linear sequence of phones, i.e., “beads-on-a-string” (Greenberg 1998; Ostendorf 1999; Young 2001; Strik 2001).

[5] `http://www.icsi.berkeley.edu/real/stp/`

In spite of the consensus that speech cannot properly be described as a linear sequence of phones, clear-cut alternatives to the "beads-on-a-string" approach do not exist. Greenberg (1998) advocates carrying out experiments according to the principles of the hypothetico-deductive method, in order to thus find out empirically what the basic "building blocks" of speech are, and how the linguistic elements are bound together to form speech. Greenberg further argues for a multi-tiered representation of speech in which only partial information from each of many levels of linguistic abstraction is required for sufficient identification of lexical and phrasal elements.

Ostendorf (1999) argues for "a finer-grained low-level representation, incorporating dependence on syllable (and higher level) structure via context conditioning." Her conclusion is that in order to move away from the beads-on-a-string model it will not suffice to simply perform pronunciation modeling or to alter the type or size of the units, but that a combination of changes to the pronunciation model and the acoustic model are needed.

What then are the implications for pronunciation variation modeling research? Should we be using syllables instead of phones? The advantages of this unit for pronunciation modeling are quite conclusively argued for in Greenberg (1999). Several researchers have since carried out experiments in which syllable structure is an integral part of their approach. Improvements in the order of 1% are reported for Switchboard by Ganapathiraju et al. (2001), in which a combination of syllables and phone models was used. On a much smaller task (OGI Alphadigits) a 20% relative performance improvement is found over a triphone system. In Wu (1998), half-syllable units were used and it was shown that incorporating syllables into an ASR system can improve continuous speech recognition accuracy and robustness for a small vocabulary corpus. However, although syllable structure is incorporated into the methods, the "beads-on-a-string" paradigm is still being employed and the improvements are comparable to what is found when modeling pronunciation variation. Therefore, is seems there is no real advantage to simply replacing phones by syllables.

Is a finer-grained, low-level representation perhaps the solution? If one looks at finer-grained representations such as articulatory-acoustic features, what are the benefits? In our work (Chang et al. 2001; Wester et al. 2001), we showed that it is possible to obtain an accurate frame-level automatic phonetic annotation without recourse to a word-level transcript using articulatory-acoustic features. However, when one attempts to convert the articulatory-acoustic features into phonetic segments, the results are not much better than a conventional phone-recognition system. This echoes results reported in Kirchhoff (1999) and King et al. (2000). Further research will have to prove whether articulatory-acoustic features can be incorporated into speech recognition systems in such a way that the benefits of these features can be exploited to obtain lower WERs.

1.8 General conclusions

In the summaries presented in Section 1.6, conclusions for each of the studies presented in this thesis were given. In this section, more general conclusions are drawn. Previous to the work presented in this thesis, Dutch pronunciation modeling for ASR was an issue that had

not yet been addressed. This thesis shows that methods developed and tested on English transfer to Dutch data.

The main goal of the research presented in this thesis was to improve the performance of Dutch ASR. Statistically significant improvements in WER were found, both for the knowledge-based and data-derived approaches (Kessens et al. 1999a; Wester 2001). The results presented in publication 1 and 3 show that in order to obtain significant improvements in WERs, prior probabilities for the variants should be incorporated in the recognition process in addition to adding variants to the lexicon.

In publication 1, another of our objectives was formulated as follows: "Our long-term goal is to find the set of rules which is optimal for modeling pronunciation variation." It is difficult to conclude whether this goal has been reached or not. It is possible that in the course of the research carried out for this thesis the optimal set of variants for the VIOS data was found. However, if that is the case, it went unnoticed, as we implicitly assumed that performing recognition with the optimal set of variants would lead to lower WERs. In Section 1.7.1, I argued that the reason for the lack of improvement in WER is because conditional probabilities are not taken into account in a static lexicon. Therefore, it could be the case that we have the correct set of variants to describe the pronunciation variation present in the VIOS material, but that this is not reflected in the WERs because of lexical confusability.

An ancillary aim of this research was to determine whether the forced recognition technique that we used in our pronunciation variation research could also be used meaningfully, in spite of its limitations, to obtain phonetic transcriptions for linguistic research. Comparing the transcriptions produced by the forced recognition to the transcriptions produced by the listeners shows that there are significant differences between the CSR and the listeners, but also that there are significant differences between listeners. Forced alignment is an extremely useful tool in speech recognition research. However, as there is no completely error-free reference transcription, the problem remains that one cannot unconditionally conclude that the CSR is concise enough, or good enough to generate transcriptions. In essence, it depends on what one is using the transcriptions for.

In addition, a limitation of the forced recognition approach is that it requires a word transcript to perform. The need for a word transcript can be evaded by using the articulatory-acoustic feature approach that was employed in publication 4. In this approach, a transcription of the speech material is possible without needing a word-level transcript. However, in order for articulatory-acoustic based features to prove truly useful for speech recognition technology, it will be necessary to develop lexical representations and pronunciation models tuned to this level of abstraction.

1.9 Future work

In Section 1.7, lexical confusability, phone transcriptions, and the beads-on-a-string paradigm were presented as shortcomings of the segmental approach to modeling pronunciation variation. This may give the impression that there is no future for pronunciation modeling. However, the outlook for pronunciation modeling is not quite that bleak. It is my impression

that the future of pronunciation modeling should lie in employing different levels of linguistic information to predict and model the variation present in the speech material. This section gives a few examples of how this can be achieved in pronunciation modeling.

The results presented in publication 3 of this thesis show that simply adding a great deal of variants to the lexicon leads to a deterioration in WER. Therefore, prior probabilities are included in the decoding process. In Section 1.7.1, it was argued that although prior probabilities are important to include in the recognition process they do not suffice for modeling pronunciation variation and that conditional probabilities are possibly the key to reducing WERs.

Different levels of linguistic information may be useful in estimating the conditional probabilities. An example of information that can be incorporated is word probability. Jurafsky et al. (2001) shows that more probable words i.e., when a word has a high unigram $P(w_i)$, a high bigram $P(w_i|w_{i-1})$, or a high reverse bigram probability $P(w_i|w_{i+1})$ then the pronunciation of that word is likely to be shorter, it is more likely that the word will have a reduced vowel and it is more likely to have a deleted /t/ or /d/. Furthermore, it was shown that function words were strongly affected by conditional probability, while content words showed weaker effects of surrounding context but strong effects of unigram probability. This type of information can be incorporated quite easily into language models. The language model can then be employed in a second pass for decoding utterances, or for dynamically adjusting which variants in the static lexicon are activated.

Other features that may be worth exploiting are suprasegmental features such as word stress, sentence stress, position of a word within an utterance, and duration. These are all features that have been shown to influence the pronunciation of words to a large extent (Ladefoged 1975; Greenberg and Chang 2000). However, attempts at incorporating stress and other prosodic factors in the speech recognition process have not yet been very successful (van Kuijk and Boves 1999; Wang and Seneff 2001), or are still in such a preliminary phase that no conclusions can be drawn yet (Shafran and Ostendorf 2000). Before these types of features can be incorporated meaningfully into ASR it is necessary to have training data that is annotated at the prosodic level. Such annotations can then be used as a starting point to analyze which information may be beneficial to pronunciation variation modeling. For example, suprasegmental features can be used as attributes for decision trees which can then be used to generate variants, or to dynamically adjust the lexicon.

In Fosler-Lussier (1999), an attempt was made at incorporating longer-range local context effects (i.e, segmental context, speaking rate, word duration and word predictability) into pronunciation models. Although, the results presented in Fosler-Lussier (1999) are slightly disappointing, the method definitely has its merits. One of the explanations given in Fosler-Lussier (1999) as to why including extra-segmental features did not improve recognition results was that these features were not robust enough for accurate prediction of pronunciation probabilities in an automatic learning system (Fosler-Lussier 1999, p. 150). This is the crux of the matter. It is of the utmost importance, if we are to incorporate extra features into the process of pronunciation modeling, that these features are robust. Therefore, finding methods of robust estimation of, for example, speaking rate and word predictability, must

also be included in future research within the field of pronunciation modeling.

To summarize, human listeners rely on many different linguistic tiers which are all used to interpret the speech signal during the course of a conversation, whereas current ASR systems use information only from a limited number of different linguistic tiers. I am convinced the future of pronunciation modeling lies in employing information from more linguistic tiers than currently are being used. Finding the correct types of information that can be exploited within the stochastic frameworks of ASR systems, and combining them in the correct way are the main hurdles that must be overcome in order to progress in ASR.

References

Amdall, I., F. Korkmazskiy, and A. Surendran (2000). Joint pronunciation modelling of non-native speakers using data-driven methods. In *Proc. of ICSLP '00*, Volume III, Beijing, pp. 622–625.

Baayen, H. (1991). De CELEX lexicale databank. *Forum der Letteren 32*(3), 221–231.

Bacchiani, M. and M. Ostendorf (1999). Joint lexicon, acoustic unit inventory and model design. *Speech Communication 29*, 99–114.

Beulen, K., S. Ortmanns, A. Eiden, S. Wartin, L. Welling, J. Overmann, and H. Ney (1998). Pronunciation modelling in the RWTH large vocabulary speech recognizer. In *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, pp. 13–16.

Biemans, M. (2000). *Gender Variation in Voice Quality*. Ph. D. thesis, University of Nijmegen.

Blackburn, C. and S. Young (1995). Towards improved speech recognition using a speech production model. In *Proc. of EUROSPEECH '95*, Madrid, pp. 1623–1626.

Booij, G. (1995). *The phonology of Dutch*. Oxford: Clarendon Press.

Bourlard, H. and N. Morgan (1993). *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers.

Brieussel-Pousse, L. and G. Perennou (1999). Language model level vs. lexical level for modeling pronunciation variation in a French CSR. In *Proc. of EUROSPEECH '99*, Budapest, pp. 1771–1774.

Chang, S., S. Greenberg, and M. Wester (2001). An elitist approach to articulatory-acoustic feature classification. In *Proc. of EUROSPEECH '01*, Aalborg, pp. 1729–1733.

Chang, S., L. Shastri, and S. Greenberg (2000). Automatic phonetic transcription of spontaneous speech (American English). In *Proc. of ICSLP '00*, Volume IV, Beijing, pp. 330–333.

Cohen, M. (1989). *Phonological Structures for Speech Recognition*. Ph. D. thesis, University of California, Berkeley, CA.

Cremelie, N. and J.-P. Martens (1999). In search of better pronunciation models for speech recognition. *Speech Communication 29*, 115–136.

Cucchiarini, C. (1993). *Phonetic Transcription: A Methodological and Empirical Study*. Ph. D. thesis, University of Nijmegen.

Deng, L. and D. Sun (1994). A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of the Acoustic Society of America 95*(5), 2702–2720.

Eide, E. (1999). Automatic modeling of pronunciation variations. In *Proc. of EUROSPEECH '99*, Budapest, pp. 451–454.

Fosler-Lussier, E. (1999). *Dynamic Pronunciation Models for Automatic Speech Recognition*. Ph. D. thesis, University of California, Berkeley, CA.

Fosler-Lussier, E. and N. Morgan (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication 29*, 137–158.

Fosler-Lussier, E. and G. Williams (1999). Not just what, but also when: Guided automatic pronunciation modeling for Broadcast News. In *DARPA Broadcast News Workshop*, Herndon, VA., pp. 171–174.

Ganapathiraju, A., J. Hamaker, M. Ordowski, G. Doddington, and J. Picone (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing 9*(4), 358–366.

Giachin, E., A. Rosenberg, and C.-H. Lee (1991). Word juncture modeling using phonological rules for HMM-based continuous speech recognition. *Computer Speech and Language 5*, 155–168.

Greenberg, S. (1998). Recognition in a new key - towards a science of spoken language. In *Proc. of ICASSP '98*, Seattle, WA., pp. 1041–1045.

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication 29*, 159–176.

Greenberg, S. and S. Chang (2000). Linguistic dissection of Switchboard-corpus automatic speech recognition systems. In *Proc. of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium*, Paris, pp. 195–202.

Greenberg, S. and E. Fosler-Lussier (2000). The uninvited guest: Information's role in guiding the production of spontaneous speech. In *Proc. of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) analysis of speech. *Journal of the Acoustic Society of America 87(4)*, 1738–1752.

Holter, T. and T. Svendsen (1999). Maximum likelihood modeling of pronunciation variation. *Speech Communication 29*, 177 –191.

Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustic Society of America 93*(1), 510–524.

Jurafsky, D., A. Bell, M. Gregory, and W. Raymond (2001). The effect of language model probability on pronunciation reduction. In *Proc. of ICASSP '01*, Salt Lake City, UT., pp. 801–804.

Kerkhoff, J. and T. Rietveld (1994). Prosody in NIROS with FONPARS and ALFEIOS. In P. de Haan and N. Oostdijk (Eds.), *Proc. of the Dept. of Language and Speech, University of Nijmegen*, Volume 18, pp. 107–119.

Kessens, J., C. Cucchiarini, and H. Strik (2001). A data-driven method for modeling pronunciation variation. *Submitted to Speech Communication*.

Kessens, J., M. Wester, C. Cucchiarini, and H. Strik (1998). The selection of pronunciation variants: comparing the performance of man and machine. In *Proc. of ICSLP '98*, Sydney, pp. 2715–2718.

Kessens, J., M. Wester, and H. Strik (1999a). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication 29*, 193–207.

Kessens, J., M. Wester, and H. Strik (1999b). Modeling within-word and cross-word pronunciation variation to improve the performance of a Dutch CSR. In *Proc. of ICPhS '99*, San Francisco, pp. 1665–1668.

King, S. and P. Taylor (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language 14*(4), 333–353.

King, S., P. Taylor, J. Frankel, and K. Richmond (2000). Speech recognition via phonetically-featured syllables. In *PHONUS 5: Proc. of the Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Institute of Phonetics, University of the Saarland, pp. 15–34.

Kipp, A., M.-B. Wesenick, and F. Schiel (1996). Automatic detection and segmentation of pronunciation variants in German speech corpora. In *Proc. of ICSLP '96*, Philadelphia, PA., pp. 106–109.

Kipp, A., M.-B. Wesenick, and F. Schiel (1997). Pronunciation modeling applied to automatic segmentation of spontaneous speech. In *Proc. of EUROSPEECH '97*, Rhodes, pp. 1023–1026.

Kirchhoff, K. (1999). *Robust Speech Recognition Using Articulatory Information*. Ph. D. thesis, University of Bielefeld.

Kirchhoff, K. (2000). Integrating articulatory features into acoustic models for speech recognition. In *PHONUS 5: Proc. of the Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Institute of Phonetics, University of the Saarland, pp. 73–86.

Ladefoged, P. (1975). *A Course in Phonetics*. Harcourt Brace Jovanovich, New York, Chicago, San Francisco, Atlanta.

Lamel, L. and G. Adda (1996). On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *Proc. of ICSLP '96*, Philadelphia, PA., pp. 6–9.

Lamel, L., R. Kassel, and S. Seneff (1986). Speech database development: Design and analysis of the acoustic-phonetic corpus. In *DARPA Speech Recognition Workshop*, pp. 100–109.

Laver, J. (1968). Voice quality and indexical information. *British Journal of Disorders of Communication 3*, 43–54.

Lee, K.-T. and C. Wellekens (2001). Dynamic lexicon using phonetic features. In *Proc. of EUROSPEECH '01*, Aalborg, pp. 1413–1416.

Ma, K., G. Zavaliagkos, and R. Iyer (1998). Pronunciation modeling for large vocabulary conversational speech recognition. In *Proc. of ICSLP '98*, Sydney, pp. 2455–2458.

McAllaster, D., L. Gillick, F. Scattone, and M. Newman (1998). Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In *Proc. of ICSLP '98*, Sydney, pp. 1847–1850.

Ostendorf, M. (1999). Moving beyond the 'beads-on-a-string' model of speech. In *Proc. of IEEE ASRU Workshop*, Keystone, CO., pp. 79–84.

Picone, J., S. Pike, T. Kamm, J. Bridle, L. Deng, Z. Ma, H. Richards, and M. Schuster (1999). Initial evaluation of hidden dynamic models on conversational speech. In *Proc. of ICASSP '99*, Phoenix, AZ., pp. 109–112.

Polzin, T. and A. Waibel (1998). Pronunciation variations in emotional speech. In *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, pp. 103–107.

Quazza, S. and H. van den Heuvel (2000). The use of lexicons in text-to-speech-systems. In F. van Eynde and D. Gibbon (Eds.), *Lexicon Development for Speech and Language Processing*, Chapter 7, pp. 207–233. Kluwer Academic Publishers.

Rabiner, L. and B.-H. Juang (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Richards, H. and J. Bridle (1999). The HDM: a segmental hidden dynamic model of coarticulation. In *Proc. of ICASSP '99*, Phoenix, AZ., pp. 357–360.

Riley, M., W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraçlar, C. Wooters, and G. Zavaliagkos (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication 29*, 209–224.

Riley, M. and A. Ljolje (1996). Automatic generation of detailed pronunciation lexicons. In C.-H. Lee, F. Soong, and K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition: Advanced Topics*, Chapter 12, pp. 285–302. Kluwer Academic Publishers.

Roach, P. and S. Arnfield (1998). Variation information in pronunciation dictionaries. In *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, pp. 121–124.

Robinson, A., G. Cook, D. Ellis, E. Fosler-Lussier, S. Renals, and D. Williams (2001). Connectionist speech recognition of Broadcast News. *To appear in Speech Communication*.

Saraçlar, M. (2000). *Pronunciation Modeling for Conversational Speech Recognition*. Ph. D. thesis, Johns Hopkins University, Baltimore, MD.

Saraçlar, M. and S. Khudanpur (2000). Pronunciation ambiguity vs. pronunciation variability in speech recognition. In *Proc. ICASSP '00*, Istanbul, pp. 1679–1682.

Saraçlar, M., H. Nock, and S. Khudanpur (2000). Pronunciation modeling by sharing Gaussian densities across phonetic models. *Computer Speech and Language 14*, 137–160.

Shafran, I. and M. Ostendorf (2000). Use of higher level linguistic structure in acoustic modeling for speech recognition. In *Proc. of ICASPP '00*, Istanbul, pp. 1021–1024.

Shriberg, L. and L. Lof (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics 5*, 225–279.

Sloboda, T. and A. Waibel (1996). Dictionary learning for spontaneous speech recognition. In *Proc. of ICSLP '96*, Philadelphia, PA., pp. 2328–2331.

Steinbiss, V., H. Ney, R. Haeb-Umbach, B.-H. Tran, U. Essen, R. Kneser, M. Oerder, H.-G. Meier, X. Aubert, C. Dugast, and D. Geller (1993). The Philips research system for large-vocabulary continuous-speech recognition. In *Proc. of EUROSPEECH '93*, Berlin, pp. 2125–2128.

Strik, H. (2001). Pronunciation adaptation at the lexical level. In *Proc. of the ITRW Adaptation Methods for Speech Recognition*, Sophia-Antipolis, pp. 123–130.

Strik, H. and C. Cucchiarini (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication 29*, 225–246.

Strik, H., A. Russel, H. van den Heuvel, C. Cucchiarini, and L. Boves (1997). A spoken dialogue system for the Dutch public transport information service. *International Journal of Speech Technology 2*(2), 119–129.

Torre, D., L. Villarrubia, L. Hernández, and J. Elvira (1996). Automatic alternative transcription generation and vocabulary selection for flexible word recognizers. In *Proc. of ICASSP '96*, Munich, pp. 1463–1466.

van Kuijk, D. and L. Boves (1999). Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication 27*, 95–111.

Wang, C. and S. Seneff (2001). Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain. In *Proc. of EUROSPEECH '01*, Aalborg, pp. 2761–2764.

Weintraub, M., K. Taussig, K. Hunicke-Smith, and A. Snodgras (1996). Effect of speaking style on LVCSR performance. In *Proc. of ICSLP '96*, Philadelphia, PA., pp. 16–19. Addendum.

Wester, M. (2001). Pronunciation modeling for ASR – knowledge-based and data-derived methods. *Submitted to Computer Speech and Language*.

Wester, M. and E. Fosler-Lussier (2000). A comparison of data-derived and knowledge-based modeling of pronunciation variation. In *Proc. of ICSLP '00*, Volume I, Beijing, pp. 270–273.

Wester, M., S. Greenberg, and S. Chang (2001). A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In *Proc. of EUROSPEECH '01*, Aalborg, pp. 1729–1732.

Wester, M., J. M. Kessens, C. Cucchiarini, and H. Strik (2001). Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. *To appear in Language and Speech 44(3)*, 377–403.

Williams, G. and S. Renals (1998). Confidence measures for evaluating pronunciation models. In *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, pp. 151–155.

Wrench, A. (2000). A multi-channel/multi-speaker articulatory database for continuous speech recognition research. In *PHONUS 5: Proc. of the Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Institute of Phonetics, University of the Saarland, pp. 1–13.

Wu, S.-L. (1998). *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*. Ph. D. thesis, University of California, Berkeley, CA.

Yang, Q. and J.-P. Martens (2000). On the importance of exception and cross-word rules for the data-driven creation of lexica for ASR. In *Proc. of the 11^{th} ProRisc Workshop*, Veldhoven, The Netherlands, pp. 589–593.

Young, S. (2001). Statistical modelling in continuous speech recognition (CSR). In *UAI '01: Proc. of the 17^{th} International Conference on Uncertainty in Artificial Intelligence*, Seattle, WA.

Appendix A

Phone symbols used in Dutch ASR

Table A.1 gives the set of SAMPA symbols that was used in the Dutch ASR systems described in this thesis. The set is based on the set listed at `http://www.phon.ucl.ac.uk/home/sampa/dutch.htm`. The corresponding IPA transciptions are also shown in Table A.1. The IPA transcription is the most likely match; in practice the SAMPA symbols encompass more than the one-to-one translation shown in Table A.1.

A few minor differences can be observed between the online SAMPA list and the set shown in Table A.1. Two of the symbols listed in Table A.1 in the column entiteld SAMPA do not occur in the online list, i.e. /L/ and /R/. These symbols were added to our set in order to enable the distinction between liquids in pre- and postvocalic position. Furthermore, a number of the symbols that occur in the online SAMPA set have not been used in this set. The reason for this is that the phones in question do not occur frequently enough to warrant training a specific model for them. Table A.2 lists these phones and their pertinent mapping.

In addition to the 37 phone models shown in Table A.1, a model for silence and a model for noise were also employed in the ASR systems.

Table A.1: SAMPA phone symbols used for ASR, their corresponding IPA transcriptions and examples of Dutch words in which the sound occurs. Relevant sound is in bold type.

#	SAMPA	IPA	Example
	Plosives		
1	p	p	**p**ak
2	b	b	**b**ak
3	t	t	**t**ak
4	d	d	**d**ak
5	k	k	**k**ap
	Fricatives		
6	f	f	**f**el
7	v	v	**v**el
8	s	s	**s**ein
9	z	z	**z**ijn
10	x	x	to**ch**
11	h	h	**h**and
12	S	ʃ	**sh**ow
	Nasals, liquids and glides		
13	m	m	**m**et
14	n	n	**n**et
15	N	ŋ	ba**ng**
16	l	l	**l**and
17	L	ɫ	ha**l**
18	r	r	**r**and
19	R	ɹ	to**r**
20	w	w	**w**it
21	j	j	**j**a

#	SAMPA	IPA	Example
	Vowels		
22	I	ɪ	p**i**t
23	E	ɛ	p**e**t
24	A	ɑ	p**a**t
25	O	ɔ	p**o**t
26	Y	œ	p**u**t
27	@	ə	g**e**mak
28	i	i	v**ie**r
29	y	y	v**uu**r
30	u	u	v**oe**r
31	a:	a	n**aa**m
32	e:	e	v**ee**r
33	2:	ø	d**eu**r
34	o:	o	v**oo**r
35	Ei	ɛi	f**ij**n
36	9y	ʌy	h**ui**s
37	Au	ɑu	g**ou**d

Table A.2: Mapped SAMPA phones.

SAMPA	IPA	Example	Mapping
g	ɡ	**g**oal	x
G	ɣ	**g**oed	x
Z	ʒ	baga**g**e	S

Part II

Publications

List of publications

This thesis consists of the following publications:

1. J.M. Kessens, M. Wester and H. Strik (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication 29*, 193-207.

2. M. Wester, J.M. Kessens, C. Cucchiarini and H. Strik (2001). Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. *Language and Speech 44(3)*, 377-403.

3. M. Wester (2001). Pronunciation modeling for ASR – knowledge-based and data-derived methods. *Submitted to Computer Speech and Language*.

4. M. Wester, S. Greenberg and S. Chang (2001). A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In *Proceedings of the 7^{th} European Conference on Speech Communication and Technology (Eurospeech-2001)*, Aalborg, pp. 1729-1732.

Other publications not included in this thesis (M. Wester first author):

1. M. Wester and E. Fosler-Lussier (2000). A comparison of data-derived and knowlegde-based modeling of pronunciation variation. In *Proceedings International Conference on Spoken Language Processing*,Volume I, Beijing, pp. 270-273.

2. M. Wester, J.M. Kessens and H. Strik (2000). Pronunciation variation in ASR: which variation to model? In *Proceedings International Conference on Spoken Language Processing*, Volume IV, Beijing, pp. 488-491.

3. M. Wester, J.M. Kessens and H. Strik (2000). Using Dutch phonological rules to model pronunciation variation in ASR. In *PHONUS 5: Proceedings of the Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Institute of Phonetics, University of the Saarland, pp. 105-116.

4. M. Wester and J.M. Kessens (1999). Comparison between expert listeners and continuous speech recognizers in selecting pronunciation variants. In *Proceedings of the 14^{th} International Congress of Phonetic Sciences*, San Francisco, pp. 723-726.

5. M. Wester, J.M. Kessens and H. Strik (1998). Two automatic approaches for analyzing the frequency of connected speech processes in Dutch. In *Proceedings International Conference on Spoken Language Processing*, Volume 7, Sydney, pp. 3351-3356.

6. M. Wester, J.M. Kessens and H. Strik (1998). Modeling pronunciation variation for a Dutch CSR: testing three methods. *Proceedings International Conference on Spoken Language Processing*, Volume 6, Sydney, pp. 2535-2538.

7. M. Wester, J.M. Kessens and H. Strik (1998). Improving the performance of a Dutch CSR by modeling pronunciation variation. *Proceedings of the ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition"*, Kerkrade, pp. 145-150.

8. M. Wester (1998). Automatic classification of voice quality: comparing regression models and hidden Markov models. In *Proceedings Voicedata98 Symposium on Databases in Voice Quality Research and Education*, Utrecht, pp. 92-97.

9. M. Wester, J.M. Kessens, C. Cucchiarini and H. Strik (1998). Selection of pronunciation variants in spontaneous speech: comparing the performance of man and machine. In *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech: Production and Perception*, Aix-en-Provence, pp. 157-160.

10. M. Wester, J.M. Kessens, C. Cucchiarini and H. Strik (1997). Modelling pronunciation variation: some preliminary results. In: H. Strik, N. Oostdijk, C. Cucchiarini and P.A. Coppen (eds.) *Proceedings of the Dept. of Language & Speech*, Vol. 20, pp. 127-137.

Other publications not included in this thesis (M. Wester co-author):

1. S. Chang, S. Greenberg and M. Wester (2001). An elitist approach to articulatory-acoustic feature classification. In *Proceedings of the 7^{th} European Conference on Speech Communication and Technology (Eurospeech-2001)*, Aalborg, pp. 1725-1728.

2. J.M. Kessens, M. Wester and H. Strik (2000). Automatic detection and verification of Dutch phonological rules. In *PHONUS 5: Proceedings of the Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Institute of Phonetics, University of the Saarland, pp. 117-128.

3. J.M. Kessens, M. Wester and H. Strik (1999). Modeling within-word and cross-word pronunciation variation to improve the performance of a Dutch CSR. In *Proceedings of the 14^{th} International Congress of Phonetic Sciences*, San Francisco, pp. 1665-1668.

4. J.M. Kessens, M. Wester, C. Cucchiarini and H. Strik (1998). The selection of pronunciation variants: comparing the performance of man and machine. In *Proceedings International Conference on Spoken Language Processing*, Volume 6, Sydney, pp. 2715-2718.

5. J.M. Kessens and M. Wester (1997). Improving recognition performance by modelling pronunciation variation. In *Proceedings of the CLS Opening Academic Year '97-'98*, Nijmegen, pp. 1-20.

6. J.M. Kessens, M. Wester, C. Cucchiarini and H. Strik (1997). Testing a method for modelling pronunciation variation. In *Proceedings of the COST workshop: Speech Technology in the Public Telephone Network: Where are we today?*, Rhodes, pp. 37-40.

Publication 1.

J.M. Kessens, M. Wester and H. Strik (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication 29*, 193-207.

ELSEVIER

Speech Communication 29 (1999) 193–207

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation

Judith M. Kessens *, Mirjam Wester, Helmer Strik

A² RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 22 December 1998; received in revised form 2 August 1999; accepted 4 August 1999

Abstract

This article describes how the performance of a Dutch continuous speech recognizer was improved by modeling pronunciation variation. We propose a general procedure for modeling pronunciation variation. In short, it consists of adding pronunciation variants to the lexicon, retraining phone models and using language models to which the pronunciation variants have been added. First, within-word pronunciation variants were generated by applying a set of five optional phonological rules to the words in the baseline lexicon. Next, a limited number of cross-word processes were modeled, using two different methods. In the first approach, cross-word processes were modeled by directly adding the cross-word variants to the lexicon, and in the second approach this was done by using multi-words. Finally, the combination of the within-word method with the two cross-word methods was tested. The word error rate (WER) measured for the baseline system was 12.75%. Compared to the baseline, a small but statistically significant improvement of 0.68% in WER was measured for the within-word method, whereas both cross-word methods in isolation led to small, non-significant improvements. The combination of the within-word method and cross-word method 2 led to the best result: an absolute improvement of 1.12% in WER was found compared to the baseline, which is a relative improvement of 8.8% in WER.

Zusammenfassung

Dieser Artikel beschreibt, wie die Leistung eines automatischen Spracherkenners, der niederländische gesprochene Sprache erkennt, mit Hilfe der Modellierung von Aussprachevarianten verbessert wurde. Für diese Modellformung wird eine allgemeine Prozedur vorgeschlagen, die – kurz gesagt – darin besteht, dem Lexikon Aussprachevarianten hinzuzufügen, die Phonmodelle erneut einer Lernphase zu unterziehen und Sprachmodelle dabei zu verwenden, in denen die Aussprachevarianten mithineinbezogen wurden. Durch Anwendung einer Gruppe von fünf optionalen phonologischen Regeln wurden im Basislexikon zunächst Aussprachevarianten innerhalb von Wörtern generiert. Dann wurde mit Hilfe zweier Methoden eine begrenzte Anzahl von Sandhiprozessen (Prozesse auf Wordgrenzen) modelliert. Die erste bestand darin, die Sandhivarianten direkt dem Lexikon hinzuzufügen und bei der zweiten wurden Multiwörter gebraucht. Letztendlich wurden die wortinternen Ausprachevarianten mit den zwei Sandhivarianten kombiniert getestet. Die Basisleistung des Spracherkenners, d.h. ohne Anwendung des Modells der Aussprachevariation, betrug 12.75% "word error rate" (WER). Bei Anwendung der wortinternen Aussprachevarianten wurde eine geringe, aber statistisch signifikante Verbesserung von 0.68% WER gemessen. Die Anwendung der zwei Sandhimodelle hingegen ergab einen

* Corresponding author. Tel.: +31(0)24-3612055; fax: +31(0)24-3612907.
E-mail address: j.kessens@let.kun.nl (J.M. Kessens)

0167-6393/99/$ - see front matter
PII: S0167-6393(99)00048-5

sehr kleinen, nicht signifikanten Verbesserung. Die Kombination des wortinternen Modells mit dem zweiten Sandhimodell hingegen ergab schließlich das beste Ergebnis: eine absolute Verbesserung von 1.12% WER, was einer relativen Verbesserung von 8.8% WER entspricht.

Résumé

Cet article décrit comment les performances d'un reconnaisseur de parole continue (CSR) pour le néerlandais ont été améliorées en modelant la variation de prononciation. Nous proposons une procédure générale pour modeler cette variation. En bref, elle consiste à ajouter des variantes de prononciation au lexique et dans le ré-apprentissage des modèles de phones en utilisant des modèles de langage auxquels les variantes de prononciation ont été ajoutées. D'abord, des variantes de prononciation à l'intérieur de mot ont été produites en appliquant un ensemble de cinq règles phonologiques optionnelles aux mots dans le lexique de base. Ensuite, un nombre limité de processus entre-mots ont été modelés, en utilisant deux méthodes différentes. Dans la première approche, des processus entre-mots ont été modelés en ajoutant directement les variantes "entre-mots" au lexique, et dans la deuxième approche ceci a été fait en utilisant des "mots-multiples". En conclusion, la combinaison de la méthode qui se limite aux processus à l'intérieur de mot avec les deux méthodes "entre-mots" a été testée. La performance de base était un taux d'erreur de 12.75% mots (WER); comparée à cette performance de base, une amélioration petite mais significative de 0.68% dans WER a été obtenue avec la méthode 'à l'intérieur de mot', tandis que les deux méthodes d'entre-mots en isolation ont mené à des petites améliorations non significatives. La combinaison de la méthode "à l'intérieur de mot" avec la méthode 2 "entre-mots" a mené au meilleur résultat: une amélioration absolue de 1.12% dans le WER a été trouvée comparée à la ligne de base, qui est une amélioration relative de 8.8% dans le WER.

Keywords: Continuous speech recognition; Modeling pronunciation variation; Within-word variation; Cross-word variation

1. Introduction

The present research concerns the continuous speech recognition component of a spoken dialog system called OVIS (Strik et al., 1997). OVIS is employed to automate part of an existing Dutch public transport information service. A large number of telephone calls of the on-line version of OVIS have been recorded and are stored in a database called VIOS. The speech material consists of interactions between man and machine. The data clearly show that the manner in which people speak to OVIS varies, ranging from using hypo-articulated speech to hyper-articulated speech. As pronunciation variation degrades the performance of a continuous speech recognizer (CSR) – if it is not properly accounted for – solutions must be found to deal with this problem. We expect that by explicitly modeling pronunciation variation some of the errors introduced by the various ways in which people address the system will be corrected. Hence, our ultimate aim is to develop a method for modeling Dutch pronunciation variation which can be used to tackle the problem of pronunciation variation for Dutch CSRs.

Since the early seventies, attempts have been made to model pronunciation variation for automatic speech recognition (for an overview see (Strik and Cucchiarini, 1998)). As most speech recognizers make use of a lexicon, a much used approach to modeling pronunciation variation has been to model it at the level of the lexicon. This can be done by using rules to generate variants which are then added to the lexicon (e.g. Cohen and Mercer, 1974; Cohen, 1989; Lamel and Adda, 1996). In our research, we also adopted this approach. First, we used four phonological rules selected from Booij (1995), which describe frequently occurring within-word pronunciation variation processes (Kessens and Wester, 1997). The results of these preliminary experiments were promising and suggested that this rule-based approach is suitable for modeling pronunciation variation. Therefore, we decided to pursue this approach and for the current research another frequent rule was added: the /r/-deletion rule (Cucchiarini and van

den Heuvel, 1995). Our long-term goal is to find the set of rules which is optimal for modeling pronunciation variation.

Our experiments showed that modeling within-word pronunciation variation in the lexicon improves the CSR's performance. However, in continuous speech there is also a lot of variation which occurs over word boundaries. For modeling cross-word variation, various methods have been tested in the past (see e.g. Cremelie and Martens, 1998; Perennou and Brieussel-Pousse, 1998; Wiseman and Downey, 1998). In our previous research (Kessens and Wester, 1997), we showed that adding multi-words (i.e. sequences of words) and their variants to the lexicon can be beneficial. Therefore, we decided to retain this approach in the current research. However, we also tested a second method for modeling cross-word variation. For this method, we selected from the multi-words the set of words which are sensitive to the cross-word processes that we focus on; cliticization, reduction and contraction (Booij, 1995). Next, the variants of these words are added to the lexicon. In other words, in this approach no multi-words (or their variants) are added to the lexicon.

In this paper, we propose a general procedure for modeling pronunciation variation. This procedure affects all three levels of the CSR at which modeling can take place: i.e. the lexicon, the phone models and the language models (Strik and Cucchiarini, 1998). Table 1 shows at which levels pronunciation variation can be incorporated in the recognition process, and the different test conditions which are used to measure the effect of adding pronunciation variation. In the abbreviations used in Table 1, the first letter indicates which type of recognition lexicon was used; either a lexicon with single (S) or multiple (M) pronunciations per word. The second letter indicates whether single (S) or multiple (M) pronunciations per word were present in the corpus used for training the phone models. The third letter indicates whether the language model was based on words (S) or on the pronunciation variants of the words (M).

The general procedure is employed to test the method for modeling within-word variation, as well as the two methods for modeling cross-word variation. First of all, the three methods were tested in isolation. We were however also interested in the results obtained when combining the different methods. Therefore, we tested a combination of modeling within-word variation together with each of the methods we used to model cross-word variation.

The question which arises here is whether the trends in recognition results measured when testing different methods for modeling pronunciation variation in isolation are the same when testing them in combination. More precisely, the question is whether the sum of the effects of the methods in isolation is (almost) the same as the total effect of the combination of the methods. The answer to this question has implications for our own research and the research on modeling pronunciation variation in general. If there are no differences in results between testing methods in isolation or in combination, it would suffice to test each method in isolation. However, if this is not the case, then all combinations will have to be tested (which poses a large practical problem, because potentially numerous combinations are possible).

This issue is important when combining methods for modeling within-and cross-word variation, but the problem can also exist within one method. Above we already mentioned that our ultimate goal is to find the optimal set of rules which describe Dutch pronunciation variation appropriately. Indeed, finding an optimal set of rules is the

Table 1
The test conditions used to measure the effect modeling pronunciation variation

	Test condition	Lexicon	Phone models	Language models
Baseline	SSS	S	S	S
1	MSS	M	S	S
2	MMS	M	M	S
3	MMM	M	M	M

goal of many rule-based approaches. If each rule can be tested in isolation the way forward is quite obvious. If, however, the outcome of modeling pronunciation variation is enormously influenced by interaction between rules, the way forward is much less straightforward. That is why we decided to pay attention to this issue.

The outline of our article is as follows. In Section 2, the CSR's baseline performance and the general procedure which we used for modeling pronunciation variation are described. A detailed description of the approaches which we used to model pronunciation variation is provided. Subsequently, in Section 3, more details about the CSR and the speech material which we used for our experiments are given. The results obtained with these methods are presented in Section 4. Finally, in Section 5, we discuss the results and their implications.

2. Method

In our research, we tested a method for modeling within-word variation (Section 2.3) and two methods for modeling cross-word variation (Section 2.4). We also tested the combination of the within-word method with each of the cross-word methods (Section 2.5). For all methods, in isolation and in combination, we employed the same general procedure. This general procedure is described in Section 2.2. The starting point, our CSR's baseline performance, is described in Section 2.1.

2.1. Baseline

The starting point of our research was to measure the CSR's baseline performance. It is crucial to have a well-defined lexicon to start out with, since any improvements or deteriorations in recognition performance due to modeling pronunciation variation are measured compared to the results obtained using this lexicon. Our baseline lexicon contains one pronunciation for each word. It was automatically generated using the transcription module of the Text-to-Speech (TTS) system developed at the University of Nijmegen (Kerkhoff and Rietveld, 1994). In this transcription module, phone transcriptions of words were obtained by looking up the transcriptions in two lexica: ONOMASTICA [1] and CELEX (Baayen, 1991). A grapheme-to-phoneme converter was employed whenever a word could not be found in either of the lexica. All transcriptions were manually checked and corrected if necessary. By using this transcription module, transcriptions of the words were obtained automatically, and consistency was achieved. A further advantage of this procedure is that it can also easily be used to add transcriptions of new words to the lexicon.

The phone models were trained on the basis of a training corpus in which the baseline transcriptions were used (see Sections 3.1 and 3.2). The language models were trained on the orthographic representation of the words in the training material. The baseline performance of the CSR was measured by carrying out a recognition test using the lexicon, phone models, and language model described above (test condition: SSS).

2.2. General procedure

Our general procedure for testing methods of modeling pronunciation variation consists of three steps:

1. In the first step, the baseline lexicon is expanded by adding pronunciation variants to it, thus creating a multiple pronunciation lexicon. Using the baseline phone models, baseline language model and this multiple pronunciation lexicon a recognition test is carried out (test condition: MSS).
2. In the second step, the multiple pronunciation lexicon is used to perform a forced recognition. In this type of recognition the CSR is "forced" to choose between different pronunciation variants of a word instead of between different words. Forced recognition is imposed through the language model. For each utterance, the language model is derived on the basis of 100 000 repetitions of the same utterance. This

[1] http://www2.echo.lu/langeng/projects/onomastica/

means that it is virtually impossible for the CSR to choose other words than the ones present in the utterance. Still, a small percentage of sentences (0.4–0.5%) are incorrectly recognized. In those cases, the baseline transcriptions are retained in the corpus. In all other cases, the baseline transcriptions are replaced by the transcription of the recognized pronunciation variants. A new set of phone models is trained on the basis of the resulting corpus containing pronunciation variants. We expect that by carrying out a forced recognition, the transcriptions of the words in the training corpus will match more accurately with the spoken utterance. Consequently, the phone models trained on the basis of this corpus will be more precise. A recognition test is performed using the multiple pronunciation lexicon, the retrained phone models and the baseline language model (test condition: MMS).

3. In the third step, the language model is altered. To calculate the baseline language model the orthographic representation of the words in the training corpus is used. Because there is only one variant per word this suffices. However, when a multiple pronunciation lexicon is used during recognition and the language model is trained on the orthographic representation of the words, all variants of the same word will have equal a priori probabilities (this probability is determined by the language model). A drawback of this is that a sporadically occurring variant may have a high a priori probability because it is a variant of a frequently occurring word, whereas the variant should have a lower a priori probability on the basis of its occurrence. Consequently, the variant may be easily confused with other words in the lexicon. A way of reducing this confusability is to base the calculation of the language model on the phone transcription of the words instead of on the orthographic transcription, i.e. on the basis of the phone transcriptions of the corpus obtained through forced recognition. A recognition test is performed using this language model, the multiple pronunciation lexicon and the updated phone models (test condition: MMM).

2.3. Method for modeling within-word pronunciation variation

The general procedure, described above, was employed to model within-word pronunciation variation. Pronunciation variants were automatically generated by applying a set of optional phonological rules for Dutch to the transcriptions in the baseline lexicon. The rules were applied to all words in the lexicon wherever it was possible and in no specific order, using a script in which the rules and conditions were specified. All of the variants generated by the script were added to the baseline lexicon, thus creating a multiple pronunciation lexicon. We modeled within-word variation using five optional phonological rules concerning: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion (SAMPA [2]-notation is used throughout this article). These rules were chosen according to the following four criteria.

First, we decided to start with rules concerning those phenomena that are known to be most detrimental to CSR. Of the three possible processes, i.e. insertions, deletions and substitutions, we expect the first two to have the largest consequences for speech recognition, because they affect the number of segments present in different realizations of the same word. Therefore, using rules concerning insertions and deletions was the first criterion we adopted. The second criterion was to choose rules that are frequently applied. Frequently applied is amenable to two interpretations. On the one hand, a rule can be frequent because it is applied whenever the context for its application is met, which means that the most frequent form would probably suffice as sole transcription. On the other hand, a rule can be frequent because the context in which the rule can be applied is very frequent (even though the rule is applied e.g. only in 50% of the cases). It is this type of frequent occurrence which is interesting because in this case it is difficult to predict which variant should be taken as the baseline form. Therefore, all possible variants should probably be included in the lexicon. The third criterion (related to the previous

[2] http://www.phon.ucl.ac.uk/home/sampa/dutch.htm

one) was that the rules should be relevant to phones that are relatively frequent in Dutch, since rules that concern infrequent phones probably have fewer consequences for the recognizer's performance. Finally, we decided to start with rules that have been extensively described in the literature, so as to avoid possible effects of overgeneration and undergeneration due to incorrect specification of the rules.

The description of the four rules: /n/-deletion, /t/-deletion, /@/-deletion and /@/-insertion is according to Booij (1995), and the description of the /r/-deletion rule is according to Cucchiarini and van den Heuvel (1995). The descriptions given here are not exhaustive, but describe how we implemented the rules.

(1) /n/-deletion: In standard Dutch, syllable-final /n/ can be dropped after a schwa, except if that syllable is a verbal stem or if it is the indefinite article *een* /@n/ "a". For many speakers, in particular in the western part of the Netherlands, the deletion of /n/ is obligatory. For example:

reizen /rEiz@n/ → /rEiz@/

(2) /r/-deletion: The rule for /r/-deletion can be divided into three parts based on the type of vowel preceding the /r/. First, /r/-deletion may occur if it is in the coda, preceded by a schwa and followed by a consonant. For example:

Amsterdam /Amst@rdAm/ → /Amst@dAm/

Second, for the cases where /r/ follows a short vowel, Cucchiarini and van den Heuvel (1995) make a distinction between unstressed and stressed short vowels. They state that after a short, stressed vowel in coda position, /r/-weakening can take place, but /r/-deletion is not allowed. However, we decided to treat /r/-weakening in the same way as /r/-deletion because there is no intermediate phone model in our phone set which describes /r/-weakening. Thus, we created pronunciation variants which, based on the rules, might be improbable, but we decided to give the CSR the possibility to choose. For example:

stressed: *Arnhem* /ARnEm/ → /AnEm/

unstressed: *Leeuwarden*
/le:wARd@n/ → /le:wAd@n/

Third, /r/-deletion may occur if it is in the coda, preceded by a long vowel and followed by a consonant. For example:

Haarlem /ha:RlEm/ → /ha:lEm/

(3) /t/-deletion: The process of /t/-deletion is one of the processes that typically occurs in fast speech, but to a lesser extent in careful speech. If a /t/ in a coda is preceded by an obstruent, and followed by another consonant, the /t/ may be deleted. For example:

rechtstreeks /rExtstre:ks/ → /rExstre:ks/

If the preceding consonant is a sonorant, /t/-deletion is possible, but then the following consonant must be an obstruent (unless the obstruent is a /k/). For example:

's avonds /sa:vOnts/ → /sa:vOns/

Although Booij does not mention that in some regional variants /t/-deletion also occurs in word-final position, we decided to apply the /t/-deletion rule in word-final position following an obstruent (unless the obstruent is an /s/). For example:

Utrecht /ytrExt/ → /ytrEx/

(4) /@/-deletion: When a Dutch word has two consecutive syllables headed by a schwa, the first schwa may be deleted, provided that the resulting onset consonant cluster consists of an obstruent followed by a liquid. For example:

latere /la:t@r@/ → /la:tr@/

(5) /@/-insertion: In nonhomorganic consonant clusters in coda position schwa may be inserted. If the second of the two consonants involved is an /s/ or a /t/, or if the cluster is a nasal followed by a homorganic consonant, /@/-insertion is not possible. Example:

Delft /dELft/ → /dEl@ft/

Each of the rules described above was tested in isolation by adding the variants to the lexicon and carrying out a recognition test. Tests were also carried out for all five rules together. In this case, all the steps of the general procedure were carried out.

2.4. Modeling cross-word pronunciation variation

The two different methods we used to model cross-word pronunciation variation are explained below. The type of cross-word variation which we modeled concerns processes of cliticization, contraction and reduction (Booij, 1995).

2.4.1. Method 1 for modeling cross-word pronunciation variation

The first step in cross-word method 1 consisted of selecting the 50 most frequently occurring word sequences from our training material. Next, from those 50 word sequences we chose those words which are sensitive to the cross-word processes cliticization, contraction and reduction. This led to the selection of seven words which made up 9% of all the words in the training corpus (see Table 2). The variants of these words were added to the lexicon and the rest of the steps of the general procedure were carried out (see Section 2.2). Table 2 shows the selected words (column 1), the total number of times the word occurs in the training material (column 2), their baseline transcriptions (column 3) and their added cross-word variants (column 4).

2.4.2. Method 2 for modeling cross-word pronunciation variation

The second method which we adopted for modeling cross-word variation was to make use of multi-words. Multi-words are word sequences which are joined together and added as separate entities to the lexicon. In order to be able to compare the results of this method to the results of the previous one, the same cross-word processes were modeled in both methods. On the basis of the seven words from cross-word method 1, multi-words were selected from the list of 50 word sequences. Only those word sequences in which at least one of the seven words was present could be chosen. Thus, 22 multi-words were selected. Subsequently, these multi-words were added to the lexicon and the language model. It was necessary for us to also add the multi-words to the language model, because effectively, for our CSR they are "new" words. Next, the cross-word variants of the multi-words were also added to the lexicon, and the remaining steps of the general procedure were carried out (see Section 2.2).

All of the selected multi-words have at least two pronunciations. If the parts of the multi-words are counted as separate words, the total number of words which could have a pronunciation variant covers 6% of the total number of words in the training corpus. This percentage is lower than that for cross-word method 1 due to the contextual constraints imposed by the multi-words. Table 3 shows the multi-words (column 1), the total number of times the multi-word occurs in the training material (column 2), their baseline transcriptions (column 3) and their added cross-word variants (column 4).

Table 2
The words selected for cross-word method 1, their counts in the training material, baseline transcriptions and added cross-word variants

Selected word	Count	Baseline	Variant(s)
ik	3578	Ik	k
dat	1207	dAt	dA
niet	1145	nit	ni
is	643	Is	s
de	415	d@	d
het	382	@t	hEt, t
dit	141	dIt	dI

2.5. Combination of the within-word and cross-word methods

In addition to testing the within-word method and the two cross-word methods in isolation, we also employed the general procedure to test the combination of the within-word method and cross-word method 1, and the combination of the within-word method and cross-word method 2. In these experiments the within-word pronunciation variants and the cross-word pronunciation variants were added to the lexica simultaneously.

For the combination of the within-word method with cross-word method 2, an extra set of experiments was carried out. This was necessary in order to be able to split the effect of adding multi-words from the effect of adding the multi-words' pronunciation variants. To achieve this, the experiments for the within-word method were repeated with the multi-words added to the lexica.

Table 3
The multi-words selected for cross-word method 2, their counts in the training material, baseline transcriptions and added cross-word variants

Multi-word	Count	Baseline	Variant(s)
ik_wil	2782	IkwIl	kwIl
dat_is	345	dAtIs	dAIs, dAs
ja_dat_klopt	228	ja:dAtklOpt	ja:dAklOpt
niet_nodig	224	nitno:d@x	nino:d@x
wil_ik	196	wIlIk	wIlk
dat_hoeft_niet	181	dAthuftnit	dAhuftnit, dAhuftni, dAthuftni
ik_heb	164	IkhEp	khEp
niet_naar	122	nitna:R	nina:R
het_is	74	@tIs	hEtIs, tIs
dit_is	74	dItIs	dIIs, dIs
niet_vanuit	72	nitvAn9yt	nivAn9yt
de_eerste	45	d@e:Rst@	de:Rst@
ik_zou	40	IkzAu	kzAu
ik_weet	38	Ikwe:t	kwe:t
ik_wilde	35	IkwIld@	kwIld@
niet_meer	31	nitme:R	nime:R
ik_hoef	31	Ikhuf	khuf
ik_moet	26	Ikmut	kmut
dit_was	25	dItwAs	dIwAs
ik_zei	24	IkzEi	kzEi
heb_ik	22	hEpIk	hEpk
is_het	20	Is@t	IshEt, Ist

The effect of the inclusion of multi-words in the language model and the lexica could then be measured by comparing these results to the results of the within-word method in isolation.

3. CSR and material

3.1. CSR

The main characteristics of the CSR are as follows. The input signals consist of 8 kHz, 8 bit A-law coded samples. Feature extraction is done every 10 ms for 16 ms frames. The first step in feature analysis is an FFT analysis to calculate the spectrum. In the following step, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied to the log filterband coefficients. The final processing stage is a running cepstral mean substraction. Besides 14 cepstral coefficients ($c_0 - c_{13}$), 14 delta coefficients are also used. This makes a total of 28 feature coefficients.

The CSR uses acoustic models, word-based language models (unigram and bigram) and a lexicon. The acoustic models are continuous density hidden Markov models (HMMs) with 32 Gaussians per state. The topology of the HMMs is as follows: each HMM consists of six states, three parts of two identical states, one of which can be skipped (Steinbiss et al., 1993). In total, 39 HMMs were trained. For each of the phonemes /l/ and /r/, two models were trained, because a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). For each of the other 33 phonemes context-independent models were trained. In addition, one model was trained for non-speech sounds and a model consisting of only one state was employed to model silence.

3.2. Material

Our training and test material, selected from the VIOS database (Strik et al., 1997), consisted of 25 104 utterances (81 090 words) and 6267 utter-

ances (21 106 words), respectively. Recordings with a high level of background noise were excluded.

The baseline training lexicon contains 1412 entries, which are all the words in the training material. Adding pronunciation variants generated by the five phonological rules (within-word method) increases the size of the lexicon to 2729 entries (an average of about 2 entries per word). The maximum number of variants that occurs for a single word is 16. For cross-word method 1, eight variants were added to the lexicon. For cross-word method 2, 22 multi-words and 28 variants of the multi-words were added to the lexicon.

The baseline test lexicon contains 1154 entries, which are all the words in the test corpus, plus a number of words which must be in the lexicon because they are part of the domain of the application, e.g. station names. The test corpus does not contain any out-of-vocabulary words. This is a somewhat artificial situation, but we did not want the CSR's performance to be influenced by words which could never be recognized correctly, simply because they were not present in the lexicon. Adding pronunciation variants generated by the five phonological rules (within-word method) leads to a lexicon with 2273 entries (also an average of about 2 entries per word). For cross-word methods 1 and 2, the same variants were added to the test lexicon as those which were added to the training lexicon.

4. Results

The results in this section are presented as best sentence word error rates (WER). The percentage WER is determined by

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100,$$

where S is the number of substitutions, D the number of deletions, I the number of insertions and N is the total number of words. During the scoring procedure only the orthographic representation was used. Whether or not the correct pronunciation variant was recognized was not taken into account. Furthermore, before scoring took place, the multi-words were split into the separate words they consist of. The significance of differences in WER was calculated with a t-test for comparison of means ($p = 0.05$) for independent samples.

Table 4 shows the results for modeling pronunciation variation for all methods in isolation, and the various combinations of methods. In Section 4.1, the results for the within-word method are described, and in Section 4.2, this is done for the two cross-word methods. Subsequently, the results of combining the within-word method with each of the cross-word methods are described in Section 4.3. In Section 4.4, a comparison is made between testing the methods in isolation and in combination. Finally, the overall results are presented in Section 4.5.

4.1. Modeling within-word pronunciation variation

Row 2 in Table 4 (within) shows the results of modeling within-word pronunciation variation. In column 2, the WER for the baseline condition (SSS) is given. Adding pronunciation variants to the lexicon (MSS) leads to an improvement of 0.31% in WER compared to the baseline (SSS). When, in addition, retrained phone models are

Table 4
WER for the within-word method (within), cross-word method 1 (cross 1), cross-word method 2 (cross 2), the within-word method with multi-words added to the lexicon and language model (within + multi), and the combination of the within-word method with cross-word method 1 (within + cross 1) and cross-word method 2 (within + cross 2)

	SSS	MSS	MMS	MMM
within	12.75	12.44	12.22	12.07
cross 1	12.75	13.00	12.89	12.59
cross 2	12.41*	12.74	12.99	12.45
within + multi	12.41*	12.05	11.81	11.72
within + cross 1	12.75	12.70	12.58	12.14
within + cross 2	12.41*	12.37	12.30	11.63

* Multi-words added to the lexicon and the language model.

used (MMS), a further improvement of 0.22% is found compared to the MSS condition. Finally, incorporating variants into the language model leads to an improvement of 0.15% compared to the MMS condition. In total, a significant improvement of 0.68% was found (SSS → MMM) for modeling within-word pronunciation variation.

4.2. Modeling cross-word pronunciation variation

Rows 3 (cross 1) and 4 (cross 2) in Table 4 show the results for each of the cross-word methods tested in isolation. It is important to note that the SSS condition for cross-word method 2 is different from the SSS condition for cross-word method 1. This is due to adding multi-words to the lexicon and the language model, which is indicated by an asterisk in Table 4. Adding multi-words to the lexicon and language model leads to an improvement of 0.34% (SSS → SSS*).

In contrast to the within-word method, adding variants to the lexicon leads to deteriorations of 0.25% and 0.33% WER for cross-word methods 1 and 2, respectively (SSS → MSS, SSS* → MSS). Although for cross-word method 1, part of the deterioration is eliminated when retrained phone models are used (MMS), there is still an increase of 0.14% in WER compared to the baseline (SSS). Using retrained phone models for cross-word method 2 leads to a further deterioration in WER of 0.25% (MSS → MMS). Adding pronunciation variants to the language model (MMM) leads to improvements of 0.30% and 0.54% for cross-word method 1 and 2 respectively, compared to the MMS condition.

Compared to the baseline, the total improvement is 0.16% for cross-word method 1, and 0.30% for cross-word method 2 (SSS → MMM). However, when the result of cross-word method 2 is compared to the SSS* condition (multi-words included), a deterioration of 0.04% is found (SSS* → MMM).

4.3. Modeling within-word and cross-word pronunciation variation

As was explained in Section 2.5, two processes play a role when using multi-words to model cross-word pronunciation variation, i.e., firstly, adding the multi-words and, secondly, adding variants of the multi-words. To measure the effect of only adding the multi-words (without variants), the experiments for within-word variation were repeated with the multi-words added to the lexicon and the language model. Row 5 in Table 4 (within + multi) shows the results of these experiments. The effect of the multi-words can be seen by comparing these results to the results of the within-word method (row 2 in Table 4). The comparison clearly shows that adding multi-words to the lexicon and the language model leads to improvements for all conditions. The improvements range from 0.34% to 0.41% for the different conditions.

In row 6 (within + cross 1) and row 7 (within + cross 2) of Table 4, the results of combining the within-word method with the two cross-word methods are shown. It can be seen that adding variants to the lexicon improves the CSR's performance by 0.05% and 0.04% for cross-word methods 1 and 2, respectively (SSS → MSS, SSS* → MSS). Using retrained phone models (MSS → MMM) improves the WER by another 0.12% for cross-word method 1, and 0.07% for cross-word method 2. Finally, the improvements are largest when the pronunciation variants are used in the language model too (MMM). For cross-word method 1, a further improvement of 0.44% is found compared to MMS, and for cross-word method 2, an even larger improvement of 0.67% is found.

For the combination of the within-word method with cross-word method 1, a total improvement of 0.61% is found for the test condition MMM compared to the baseline (SSS). For the same test condition, the combination of the within-word method with cross-word method 2 leads to a total improvement of 0.78% compared to the SSS* condition.

4.4. Comparing methods in isolation and in combination

In order to get a clearer picture of the differences in results obtained when modeling pronunciation variation in isolation and in combination,

the results presented in the previous sections were analyzed to a further extent.

First, the difference in WER (ΔWER) between each of the methods tested in isolation and the baseline was calculated. Next, the ΔWER for each of the cross-word methods in isolation was added to the ΔWER for the within-word method in isolation. The results of these summations are indicated by the "sum" bars in Figs. 1 and 2. The differences in WER between the baseline and the combinations of within-word and cross-word methods 1 and 2 were also calculated. These results are shown in Figs. 1 and 2 and are indicated by the "combi" bars. Fig. 1 shows the results for cross-word method 1, and Fig. 2 shows the results for cross-word method 2.

In these figures, it can be seen that the sum of the improvements for the two methods tested in isolation is not the same as the improvement obtained when testing the combinations of the methods. For cross-word method 1, the sum of the methods in isolation gives better results, whereas for cross-word method 2, the combination leads to higher improvements.

Fig. 3 shows the differences in WER between the results of adding variants of each of the five phonological rules to the lexicon separately, the summation of these results ("sum") and the result of the combination of all five rules ("combi"). The differences shown in Fig. 3 are all on the basis of the MSS condition, i.e. variants are only added to the lexicon. In isolation, the rule for /n/-deletion leads to an improvement. The variants generated by the rules for /r/-deletion and /@/-deletion seem to have almost no effect at all. The variants for /t/-deletion and /@/-insertion have some effect, but lead to a deterioration in WER compared to the baseline. The sum of these results is a deterioration

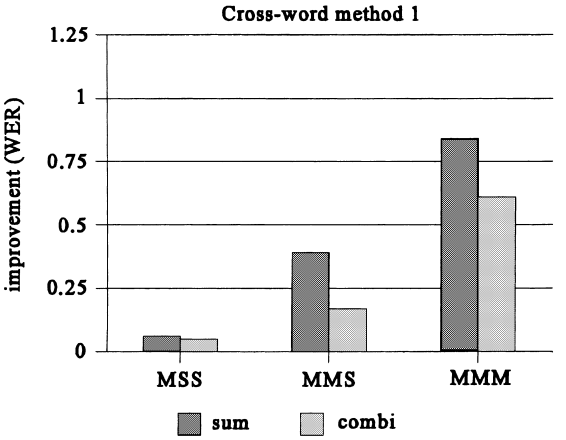


Fig. 1. Improvements (WER) for cross-word method 1 combined with the within-word method and the sum of the two methods in isolation.

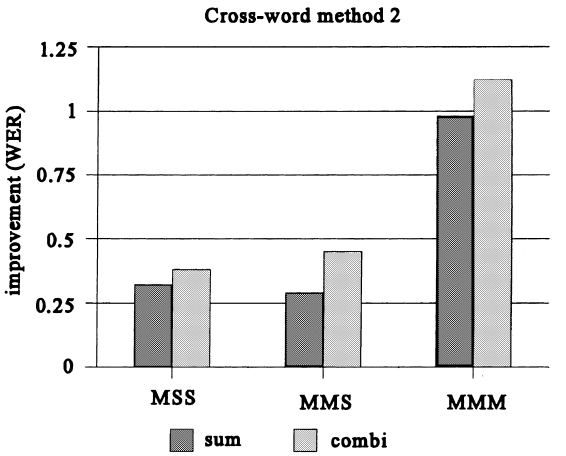


Fig. 2. Improvements (WER) for cross-word method 2 combined with the within-word method and the sum of the two methods in isolation.

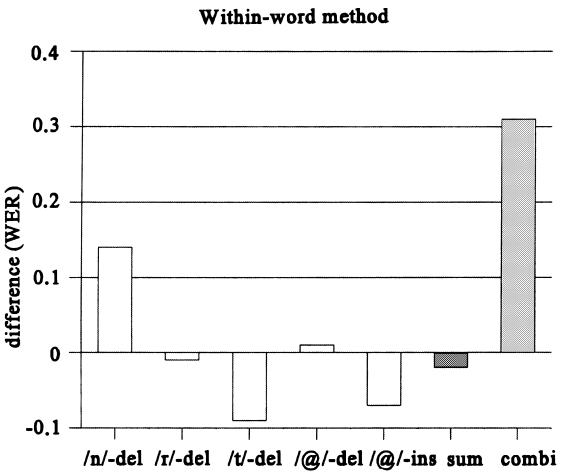


Fig. 3. Difference in WER between the baseline result and results of adding variants of separate rules to the lexicon, sum of those results, and combination result of all rules.

in WER of 0.02%. However, combining all methods, leads to an improvement of 0.31% compared to the baseline.

4.5. Overall results

For all methods, the best results are obtained when pronunciation variants are used during training and recognition, and when they are added to the language model (MMM). All methods lead to an improvement in the CSR's performance when their results are compared to the result of the baseline (SSS). These improvements are summed up in Table 5. Modeling within-word variation in isolation gives a significant improvement of 0.68%, and in combination with cross-word method 2, the improvement is also significant.

Up until now we have only presented our results in terms of WER (as is done in most studies). WERs give an indication of the net change in the performance of one CSR compared to another one. However, they do not provide more detailed information on how the recognition results of the two CSRs differ. Since this kind of detailed information is needed to gain more insight, we carried out a partial error analysis. To this end, we compared the utterances recognized with the baseline test to those recognized with our best test (MMM for within + cross 2 in Table 4). For the moment, we have restricted our error analysis to the level of the whole utterance, mainly for practical reasons. In the near future, we plan to do it at the word level too.

The results in Table 6 show how many utterances in the test corpus are actually recognized correctly or incorrectly in the two tests. These results show that 75.7% of the utterances are recognized correctly in both conditions (baseline test correct, final test correct), and 17.3% of the utterances are recognized incorrectly in both conditions. Improvements are found for 4.3% of the utterances (baseline test incorrect, final test correct), and deteriorations are found for 2.9% of the utterances (baseline test correct, final test incorrect).

The comparison of the utterances recognized differently in the two conditions can also be used to study how many changes truly occur. These results are presented in Table 7. The group of 1083 utterances (17.3%) which are recognized incorrectly in both tests (see Table 6) consist of 609 utterances (9.7%) for which both tests produce the same incorrect recognition results and 474 utterances ($17.3 - 9.7 = 7.6\%$) with different mistakes. In addition, improvements were found for 267 utterances (4.3%) and deteriorations for 183 utterances (2.9%), as was already mentioned above. Consequently, the net result is an improvement for only 84 utterances ($267 - 183$), whereas in total the recognition result changes for 924 utterances ($474 + 267 + 183$). These changes are a consequence of our methods of modeling pronunciation variation, but they cannot be seen in the WER.

Table 5
ΔWER for condition MMM compared to the baseline (SSS) for all methods

Method	ΔWER
within	0.68*
cross 1	0.16
cross 2	0.30
within + cross 1	0.61
within + cross 2	1.12*

* Significant improvements.

Table 6
Comparison between baseline test and final test condition: number of correct utterances, incorrect utterances, improvements and deteriorations (percentages between brackets)

		Baseline test	
		Correct	Incorrect
Final test	Correct	4743(75.7%)	267 (4.3%)
	Incorrect	183 (2.9%)	1083(17.3%)

Table 7
Type of change in utterances going from baseline condition to final test condition (percentages between brackets)

Type of change	Number of utterances
Same utterance, different mistake	474 (7.6%)
Improvements	267 (4.3%)
Deteriorations	183 (2.9%)
Net result	+84 (1.3%)

The WER only reflects the net result obtained, and our error analysis has shown that this is only a fraction of what actually happens due to applying our methods.

5. Discussion

In this research, we attempted to model two types of variation: within-word variation and cross-word variation. To this end, we used a general procedure in which pronunciation variation was modeled at the three different levels in the CSR: the lexicon, the phone models and the language model. We found that the best results were obtained when all of the steps of the general procedure were carried out, i.e. when pronunciation variants were incorporated at all three levels. Below, the results of incorporating pronunciation variants at all three levels are successively discussed.

In the first step, variants were only incorporated at the level of the *lexicon*. Compared to the baseline (SSS → MSS), an improvement was found for the within-word method and for the within-word method in combination with each of the two cross-word methods. However, a deterioration was found for the two cross-word methods in isolation. A possible explanation for the deterioration for cross-word method 1 is related to the fact that the pronunciation variants of cross-word method 1 are very short (see Table 2); some of them consist of only one phone. Such short variants can easily be inserted; for instance, the plosives /k/ and /t/ might occasionally be inserted at places where clicks in the signal occur. Furthermore, this effect is facilitated by the high frequency of occurrence of the words involved, i.e. they are favored by the language model. Similar things might happen for cross-word method 2. Let us give an example to illustrate this: A possible variant of the multi-word "ik_wil" /IkwIl/ is /kwIl/. The latter might occasionally be confused with the word "wil" /wIl/. This confusion leads to a substitution, but effectively it is the insertion of the phone /k/. Consequently, insertion of /k/ and other phones is also possible in cross-word method 2, and this could explain the deterioration found for cross-word method 2.

When, in the second step, pronunciation variation is also incorporated at the level of the *phone models* (MSS → MMS), the CSR's performance improved in all cases, except in the case of cross-word method 2. A possible cause of this deterioration in performance could be that the phone models were not retrained properly. During forced recognition, the option for recognizing a pause between the separate parts of the multi-words was not given. As a consequence, if a pause occurred in the acoustic signal of a multi-word, the pause was used to train the surrounding phone models, which results in contaminated phone models. Error-analysis revealed that in 5% of the cases a pause was indeed present within the multi-words in our training material. Further research will have to show whether this was the only cause of the deterioration in performance or whether there are other reasons why retraining phone models using multi-words did not lead to improvements.

In the third step, pronunciation variants were also incorporated at the level of the *language model* (MMS → MMM), which is beneficial to all methods. Moreover, the effect of adding variants to the language model is much larger for the cross-word methods than for the within-word method. This is probably due to the fact that many recognition errors introduced in the first step (see above) are corrected when variants are also included in the language model. When cross-word variants are added to the lexicon (step 1), short sequences of only one or two phones long (like e.g. the phone /k/) can easily be inserted, as was argued above. The output of forced recognition reveals that the cross-word variants occur less frequently than the canonical pronunciations present in the baseline lexicon: on average in about 13% of the cases for cross-word method 1, and 9% for cross-word method 2. In the language model with cross-word variants included, the probability of these cross-word variants is thus lower than in the original language model and, consequently, it is most likely that they will be inserted less often.

One of the questions we posed in the introduction was what the best way of modeling cross-word variation is. On the basis of our results we

can conclude that when cross-word variation is modeled in isolation, cross-word method 2 performs better than cross-word method 1, but the difference is non-significant. In combination with the within-word method, cross-word method 2 leads to an improvement compared to the within-word method in isolation. This is not the case for cross-word method 1, which leads to a degradation in WER. Therefore, it seems that cross-word method 2 is more suitable for modeling cross-word pronunciation variation. It should be noted, however, that most of the improvements gained with cross-word method 2 are due to adding the multi-words to the lexicon and the language model. An explanation for these improvements is that by adding multi-words to the language model the span of the unigram and bigram increases for the most frequent word sequences in the training corpus. Thus, more context information can be used during the recognition process. Furthermore, it should also be noted that only a small amount of data was involved in the cross-word processes which were studied; only 6–9% of the words in the training corpus were affected by these processes. Therefore, we plan to test cross-word methods 1 and 2 for a larger amount of data and a larger number of cross-word processes.

In Section 4.4, it was shown that testing the within-word method and cross-word method 2 in combination leads to better results than the sum of the results of testing the two methods in isolation. For cross-word method 1 the opposite is true, the within-word method in isolation leads to better results. The results for the within-word method show the difference which exists between testing methods in isolation or in combination even more clearly. The sum of the results for separate rules leads to a degradation in WER (compared to the baseline), whereas the combination leads to an improvement. It is clear that the principle of superposition does not apply here, neither for the five rules of the within-word method nor for the within-word method in combination with each of the two cross-word methods. This is due to a number of factors. First of all, different rules can apply to the same words. Consequently, when the five rules are used in combination, pronunciation variants are generated which are not generated for any of the rules in isolation. Furthermore, when methods are employed in combination, confusion can occur between pronunciation variants of each of the different methods. It is obvious that this confusion cannot occur when methods are tested in isolation. Finally, during decoding, the words in the utterances are not recognized independently of each other, and thus, interaction between pronunciation variants can occur. The implication of these findings is that it will not suffice to study methods in isolation. Instead, they will have to be studied in combination. However, this poses a practical problem as there are many possible combinations.

In Sections 4.1–4.4, various methods and their combinations were tested. This was done by calculating the WER after a method had been applied, and comparing this number to the WER of the baseline system. This amount of reduction in WER is a measure which is used in many studies about modeling pronunciation variation (see Strik and Cucchiarini, 1998). Although this measure gives a global idea of the merits of a method, it certainly does not reveal all details of the effect a method has. This became clear through the error analysis which we conducted (see Section 4.4). This error analysis showed that 14.7% of the recognized utterances changed, whereas a net improvement of only 1.3% in the sentence error rate was found (and 1.12% in the WER). Therefore, it is clear that a more detailed error analysis is necessary to obtain real insight into the effect of a certain method.

That is why we intend to carry out more detailed error analyses in the near future. Such a detailed error analysis should not be carried out on the test corpus, because then the test corpus is no longer an independent test set. Therefore, we will be using a development test set to do error analysis. Furthermore, instead of analyzing errors at the level of the whole utterance, we will be looking at the word level, and if necessary at the level of the phones. Through an error analysis, the effect of testing methods in isolation and in combination can be analyzed. It is hoped that this will yield the tools which are needed to decide beforehand which types of pronunciation variation should be modeled and how they should be tested.

To summarize, we obtained the best results when within-word pronunciation variation and cross-word pronunciation variation using multi-words were modeled in combination, and all the steps of the general procedure had been carried out. Using only five phonological rules and 22 multi-words a relative improvement of 8.8% was found (12.75%–11.63%).

Acknowledgements

The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research). The research by Helmer Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

References

Baayen, H., 1991. De CELEX lexicale databank. Forum der Letteren 32 (3), 221–231.

Booij, G., 1995. The Phonology of Dutch. Clarendon Press, Oxford.

Cohen, M.H., 1989. Phonological structures for speech recognition. Ph.D. dissertation. University of California, Berkeley.

Cohen, P.S., Mercer, R.L., 1974. The phonological component of an automatic speech-recognition system. In: Erman, L. (Ed.), Proceedings of the IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, 15–19 April 1974, pp. 177–187.

Cremelie, N., Martens, J.-P., 1998. In search of pronunciation rules. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 23–27.

Cucchiarini, C., van den Heuvel, H., 1995. /r/ deletion in standard Dutch. In: Strik et al. (Eds.), Proceedings of the Department of Language and Speech, University of Nijmegen, Vol. 19, pp. 59–65.

Kerkhoff, J., Rietveld, T., 1994. Prosody in Niros with Fonpars and Alfeios. In: de Haan, Oostdijk (Eds.), Proceedings of the Department of Language and Speech, University of Nijmegen, Vol. 18, pp. 107–119.

Kessens, J.M., Wester, M., 1997. Improving recognition performance by modeling pronunciation variation. In: Proceedings of the CLS opening Academic Year '97–'98, pp. 1–19. http://lands.let.kun.nl/literature/kessens.1997.1.html.

Lamel, L.F., Adda, G., 1996. On designing pronunciation lexica for large vocabulary continuous speech recognition. In: Proceedings of ICSLP-96, Philadelphia, pp. 6–9.

Perennou, G., Brieussel-Pousse, L., 1998. Phonological component in automatic speech recognition. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 91–96.

Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C., Geller, D., 1993. The philips research system for large-vocabulary continuous-speech recognition. In: Proceedings of the ESCA Third European Conference on Speech Communication and Technology: EUROSPEECH '93, Berlin, pp. 2125–2128.

Strik, H., Cucchiarini, C., 1998. Modeling pronunciation variation for ASR: Overview and comparison of methods. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 137–144.

Strik, H., Russel, A., van den Heuvel, H., Cucchiarini, C., Boves, L., 1997. A spoken dialogue system for the Dutch public transport information service. Internat. J. Speech Technol. 2 (2), 119–129.

Wiseman, R., Downey, S., 1998. Dynamic and static improvements to lexical baseforms. In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, 4–6 May 1998. A^{2}RT, University of Nijmegen, pp. 157–162.

Publication

2.

M. Wester , J.M. Kessens, C. Cucchiarini and H. Strik (2001). Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. *Language and Speech 44(3)*, 377-403.

*Obtaining Phonetic Transcriptions: A Comparison between Expert Listeners and a Continuous Speech Recognizer**

Mirjam Wester, Judith M. Kessens, Catia Cucchiarini, and Helmer Strik

University of Nijmegen

Key words

automatic transcription

continuous speech recognition

pronunciation variation

Abstract

In this article, we address the issue of using a continuous speech recognition tool to obtain phonetic or phonological representations of speech. Two experiments were carried out in which the performance of a continuous speech recognizer (CSR) was compared to the performance of expert listeners in a task of judging whether a number of prespecified phones had been realized in an utterance. In the first experiment, nine expert listeners and the CSR carried out exactly the same task: deciding whether a segment was present or not in 467 cases. In the second experiment, we expanded on the first experiment by focusing on two phonological processes: schwa-deletion and schwa-insertion. The results of these experiments show that significant differences in performance were found between the CSR and the listeners, but also between individual listeners. Although some of these differences appeared to be statistically significant, their magnitude is such that they may very well be acceptable depending on what the transcriptions are needed for. In other words, although the CSR is not infallible, it makes it possible to explore large datasets, which might outweigh the errors introduced by the mistakes the CSR makes. For these reasons, we can conclude that the CSR can be used instead of a listener to carry out this type of task: deciding whether a phone is present or not.

* *Acknowledgments:* We kindly thank Prof. Dr. W.H. Vieregge for integrating our transcription material in his course curriculum. We are grateful to the various members of A^2RT who gave their comments on previous versions of this article. We would like to thank Stephen Isard, Julia McGory, and Ann Syrdal for their useful comments on an earlier version of this article. The research by J.M. Kessens was carried out within the framework of the Priority Program Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research). The research by Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

Address for correspondence: Mirjam Wester, A^2RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands;
e-mail: <M.Wester@let.kun.nl>

1 Introduction

In the last decade, an increasing number of databases have been recorded for the purpose of speech technology research (see for instance: <http://www.ldc.upenn.edu> and <http://www.icp.inpg.fr/ELRA/>). What started out as recordings of isolated words in restricted domains has now evolved to recordings of spontaneous speech in numerous domains. Since these databases contain a wealth of information concerning human language and speech, it seems that they should somehow be made available for linguistic research in addition to the speech technology research for which they were originally constructed and are currently being employed.

The use of such databases for linguistic research has at least two important advantages. First, many of them contain spontaneous speech. Most of the knowledge on speech production and perception is based on so-called "laboratory speech," while spontaneous speech is still under-researched (Cutler, 1998; Duez, 1998; Mehta & Cutler, 1988; Rischel, 1992; Swerts & Collier, 1992). Since it is questionable whether the findings concerning laboratory speech generalize to spontaneous speech, it seems that more emphasis should be placed on studying spontaneous speech. Second, these databases contain large amounts of speech material, which bodes well for the generalizability of the results of research that uses these databases as input.

Recent studies that have made use of such large databases of spontaneous speech reveal that this line of research is worth pursuing (Greenberg, 1999; Keating, 1997). On the basis of these observations one could get the impression that analysis of the speech data contained in such databases is within the reach of any linguist. Unfortunately, this is not true. The information stored in these databases is not always represented in a way that is most suitable for linguistic research. In general, before the speech material contained in the databases can be used for linguistic research it has to be phonetically transcribed (see, for instance, Greenberg, 1999). Phonetic transcriptions are obtained by analyzing an utterance auditorily into a sequence of speech units represented by phonetic symbols and making them is therefore extremely time-consuming. For this reason, linguists often decide not to have whole utterances transcribed, but only those parts of the utterance where the phenomenon under study is expected to take place (e.g., Kuijpers & van Donselaar, 1997). In this way, the amount of material to be transcribed can be limited in a way that is least detrimental for the investigation being carried out. Nevertheless, even with this restriction, making phonetic transcriptions remains a time-consuming, costly and often tedious task.

Another problem with manual phonetic transcriptions is that they tend to contain an element of subjectivity (Amorosa, von Benda, Wagner, & Keck, 1985; Laver, 1965; Oller & Eilers, 1975; Pye, Wilcox, & Siren, 1988; Shriberg & Lof, 1991; Ting, 1970; Witting, 1962). These studies reveal that transcriptions of the same utterance may show considerable differences, either when they are made by different transcribers (between-subjects variation) or when they are made by the same transcriber, but at different times or under different conditions (within-subjects variation). Since the presence of such discrepancies throws doubt on the reliability of phonetic transcription, it has become customary among researchers who use transcription data for their studies to have more than one person transcribe the speech material (e.g., Kuijpers & van Donselaar, 1997). This of course makes the task of transcribing speech even more time-consuming and costly.

To summarize, the problems connected with obtaining good manual phonetic transcriptions impose limitations on the amount of material that can be analyzed in linguistic research, with obvious consequences for the generalizability of the results. This suggests that if it were possible to obtain good phonetic transcriptions automatically, linguistic research would be made easier. Furthermore, in this way linguistic research could make profitable use of the large speech databases.

In speech technology, various tools have been developed that go some way toward obtaining phonetic representations of speech in an automatic manner. It is possible to obtain complete unrestricted phone-level transcriptions from scratch. However, phone accuracy turns out to vary between approximately 50% and 70%. For our continuous speech recognizer, we measured a phone accuracy level of 63% (Wester, Kessens, & Strik, 1998). In general, such levels of phone accuracy are too low for many applications. Therefore, to achieve acceptable recognition results, top-down constraints are usually applied.

The top-down constraints generally used in standard CSRs are a lexicon and a language model. With these constraints, word accuracy levels are obtained which are higher than the phone accuracy levels just mentioned. However, the transcriptions obtained with standard CSRs are not suitable for linguistic research because complete words are recognized, leading to transcriptions that are not detailed enough. The transcriptions thus obtained are simply the canonical transcriptions that are present in the lexicon. More often than not, the lexicon contains only one entry for each word thus always leading to the same transcription for a word regardless of pronunciation variation, whereas for linguistic research it is precisely this detail, a phone-level transcription, which is needed.

A way of obtaining a representation that approaches phonetic transcription is by using forced recognition, also known as forced (Viterbi) alignment. In forced recognition, the CSR is constrained by only allowing it to recognize the words present in the utterance being recognized. Therefore, in order to perform forced recognition, the orthographic transcription of the utterance is needed. The forced choice entails choosing between several pronunciation variants for each of the words present in the utterance. In this way, the variants that most closely resemble what was said in an utterance can be chosen. In other words, by choosing alternative variants that differ from each other in the representation of one specific segment, the CSR can be forced, as it were, to choose between different transcriptions of that specific segment thus leading to a transcription which is more detailed than a simple word-level transcription.

A problem of automatic transcription is the evaluation of the results. Given that there is no absolute truth of the matter as to what phones a person has produced, there is also no reference transcription that can be considered correct and with which the automatic transcription can be compared (Cucchiarini, 1993, pp. 11 – 13). To try and circumvent this problem as much as possible, different procedures have been devised to obtain reference transcriptions. One possibility consists in using a consensus transcription, which is a transcription made by several transcribers after they have agreed on each individual symbol (Shriberg, Kwiatkowski, & Hoffman, 1984). Another option is to have more than one transcriber transcribe the material and to use only that part of the material for which all transcribers agree or at least the majority of them (Kuijpers & van Donselaar, 1997).

The issues of automatic transcription and its evaluation have been addressed for example, by Kipp, Wesenick, and Schiel (1997) within the framework of the Munich

Automatic Segmentation System. The performance of MAUS has been evaluated by comparing the automatically obtained transcriptions with transcriptions made by three experts. The three manual transcriptions were not used to compose a reference transcription, but were compared pairwise with each other and with the automatic transcriptions to determine the degree of agreement. The results showed that the percentage agreement ranged from 78.8% to 82.6% for the three human transcribers, while agreement between MAUS and any of the human transcriptions ranged from 74.9% to 80.3% using data-driven rules, and from 72.5% to 77.2% using rules compiled by an experienced phonetician. These results indicate how the degree of agreement differs between expert transcribers and an automatic system, and, in a sense, this is a way of showing that the machine is just one of the transcribers. However, this is not sufficient because it does not say much about the quality of the transcriptions of the individual transcribers. Therefore, we propose the use of a reference transcription.

The aim of our research is to determine whether the automatic techniques that have been developed to obtain some sort of phonetic transcriptions for CSR can also be used meaningfully, in spite of their limitations, to obtain phonetic transcriptions for linguistic research. To answer this question, we started from an analysis of the common practice in many (socio/psycho) linguistic studies in which, as mentioned above, only specific parts of the speech material have to be transcribed. In addition, we further restricted the scope of our study by limiting it to insertion and deletion phenomena, which is to say that we did not investigate substitutions. The rationale behind this choice is that it should be easier for a CSR to determine whether a segment is present or not than to determine which one of several variants of a given segment has been realized. If the technique presented here turns out to work for deletions and insertions it could then be extended to other processes. In other words, our starting point was a clear awareness of the limitations of current CSR systems, and an appreciation of the potentials that CSR techniques, despite their present limitations, could have for linguistic research.

In this study, we describe two experiments in which different comparisons are carried out between the automatically obtained transcriptions and the transcriptions made by human transcribers. In these experiments the two most common approaches to obtaining a reference transcription are used: the majority vote procedure and the consensus transcription.

In the first experiment, four kinds of comparisons are carried out to study how the machine's performance relates to that of nine listeners. First of all the degree of agreement in machine-listener pairs is compared to the degree of agreement in listener-listener pairs, as in the Kipp et al. (1997) study. Second, in order to be able to say more about the quality of the machine's transcriptions and the transcriptions by the nine listeners, they are all compared to a reference transcription (majority vote procedure). Third, because it can be expected that not all processes give the same results, the comparisons with the reference transcription are carried out for each individual process of deletion and insertion. Fourth, a more detailed comparison of the choices made by the machine and by the listeners is carried out to get a better understanding of the differences between the machine's performance and that of the listeners.

The results of this last comparison show that the CSR systematically tends to choose for deletion (non-insertion) of phones more often than listeners do. To analyze this to a further

extent, we carried out a second experiment in order to find out why and in what way the detection of a phone is different for the CSR and for the listeners. In order to study this, a more detailed reference transcription was needed. Therefore, we used a consensus transcription instead of a majority vote procedure to obtain a reference transcription.

The organization of this article is as follows: First, the methodology of the first experiment is explained followed by the presentation of the results. Before going on to the second experiment a discussion of the results of Experiment 1 is given. Following on from this, the methodology of the second experiment is explained, subsequently the results are shown and also discussed. Finally, conclusions are drawn as to the merits and usability of our automatic transcription tool.

2 Experiment 1

2.1 *Method and Material*

2.1.1 *Phonological variation*

The processes we chose to study concern insertions and deletions of phones within words (i.e., alterations in the number of segments). Five phonological processes were selected for investigation: /n/-deletion, /r/-deletion, /t/-deletion, schwa-deletion and schwa-insertion. The main reasons for selecting these five phonological processes are that they occur frequently in Dutch and are well described in the linguistic literature. Furthermore, these phonological processes typically occur in fast or extemporaneous speech, but to a lesser extent in careful speech; therefore it is to be expected that they will occur in our speech material (for more details on the speech material, see the following section).

The following description of the four processes: /n/-deletion, /t/-deletion, schwa-deletion and schwa-insertion is according to Booij (1995), and the description of the /r/-deletion process is according to Cucchiarini and van den Heuvel (1999). The descriptions given here are not exhaustive, but describe the conditions of rule application which we formulated to generate the variants of the phonological processes.

1. /n/-deletion:

In standard Dutch, syllable-final /n/ can be dropped after a schwa, except if that syllable is a verbal stem or if it is the indefinite article *een* [ən] 'a'. For many speakers, in particular in the western part of the Netherlands, the deletion of /n/ is obligatory.

Example: *reizen* [rɛizən] → [rɛizə] 'to travel'

2. /r/-deletion:

According to Cucchiarini and van den Heuvel (1999), /r/-deletion can take place in Dutch when /r/ is preceded by a vowel and followed by a consonant in a word. Although this phenomenon is attested in various contexts, it appears to be significantly more frequent when the vowel preceding the /r/ is a schwa.

Example: *Amsterdam* [ɑmstərdɑm] → [ɑmstədɑm] 'Amsterdam'

3. /t/-deletion:

If a /t/ in a coda is preceded by an obstruent, and followed by another consonant, the /t/ may be deleted.

Example: *rechtstreeks* [rɛxtstreks] → [rɛxstreks] 'directly'

If the preceding consonant is a sonorant, /t/-deletion is possible, but then the following consonant must be an obstruent (unless the obstruent is a /k/).

Example: *'s avonds* [savɔnts] → [savɔns] 'in the evening'

Finally, we also included /t/-deletion in word-final position following an obstruent.

Example: *Utrecht* [ytrɛxt] → [ytrɛx] 'Utrecht'

4. schwa-deletion:

When a Dutch word has two consecutive syllables headed by a schwa, the first schwa may be deleted, provided that the resulting onset consonant cluster consists of an obstruent followed by a liquid.

Example: *latere* [latərə] → [latrə] 'later'

5. schwa-insertion:

In nonhomorganic consonant clusters in coda position schwa may be inserted. Schwa-insertion is not possible if the second of the two consonants involved is an /s/ or a /t/, or if the cluster is a nasal followed by a homorganic consonant.

Example: *Delft* [dɛlft] → [dɛləft] 'Delft'

2.1.2
Selection of speech material

The speech material used in the experiments was selected from a Dutch database called VIOS, which contains a large number of telephone calls recorded with the on-line version of a spoken dialog system called OVIS (Strik, Russel, Van Den Heuvel, Cucchiarini, & Boves, 1997). OVIS is employed to automate part of an existing Dutch public transport information service. The speech material consists of interactions between man and machine, and can be described as extemporaneous speech.

The phonological rules described in the previous section were used to automatically generate pronunciation variants for the words being studied. In some cases, it was possible to apply more than one rule to the same word. However, in order to keep the task relatively easy for the listeners we decided to limit to two the number of rules which could apply to a single word.

From the VIOS corpus, 186 utterances were selected. These utterances contain 379 words with relevant contexts for one or two rules to apply. For 88 words, the conditions for rule application were met for two rules simultaneously and thus four pronunciation variants were generated. For the other 291 words, only one condition of rule application was relevant and two variants were generated. Consequently, the total number of instances in which a rule could be applied is 467. Table 1 shows the number of items for each of the different rules and the percentages of the total number of items. This distribution (columns 2 and 3) is not uniform, because the distribution in the VIOS corpus (columns 4 and 5) is

TABLE 1

Number of items selected per process for Experiment 1, and the percentage of the total number of items in Experiment 1. Number of items and their corresponding percentages in the VIOS corpus, for each process

phonological process	*# Exp. 1*	*% Exp. 1*	*# VIOS corpus*	*% VIOS corpus*
/n/-deletion	155	33.2	10,694	45.2
/r/-deletion	127	27.2	7,145	30.2
/t/-deletion	84	18.0	3,665	15.5
schwa-deletion	53	11.3	275	1.2
schwa-insertion	48	10.3	1,871	7.9

not uniform. However, we tried to ensure a more even distribution by having at least a 10% representation for each phonological process in the material which was selected for Experiment 1.

2.1.3 *Experimental procedure*

Nine expert listeners and the continuous speech recognizer (CSR) carried out the same task, that is, deciding for the 379 words which pronunciation variant best matched the word that had been realized in the spoken utterances (forced choice).

Listeners. The nine expert listeners are all linguists who were selected to participate in this experiment because they have all carried out similar tasks for their own investigations. For this reason, they are representative of the kind of people that make phonetic transcriptions and who may benefit from automatic ways of obtaining such transcriptions. The 186 utterances were presented to them over headphones, in three sessions, with the possibility of a short break between successive sessions. The orthographic representation of the whole utterance was shown on screen, see Figure 1. The words which had to be judged were indicated by an asterisk. Beneath the utterance, the phonemic transcriptions of the pronunciation variants were shown. The listeners' task was to indicate for each word which of the phonemic transcriptions presented best corresponded to the spoken word. The listener could listen to an utterance as often as he/she felt was necessary in order to judge which pronunciation variant had been realized.

CSR. The utterances presented to the listeners were also used as input to the CSR which is part of the spoken dialog system OVIS (Strik et al., 1997). The orthography of the utterances was available to the CSR. The main components of the CSR are a lexicon, a language model, and acoustic models.

For the automatic transcription task, the CSR was used in forced recognition mode. In this type of recognition, the CSR is "forced" to choose between different pronunciations of a word instead of between different words. Hence, a lexicon with more than one possible pronunciation per word was needed. This lexicon was made by generating pronunciation

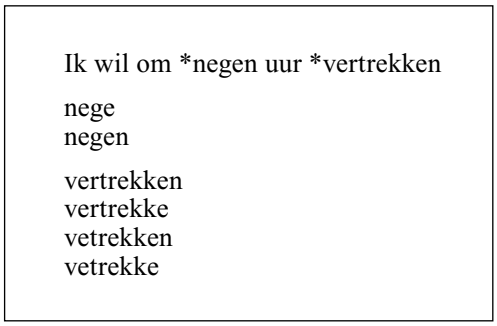


Figure 1

Pronunciation variant selection by the nine expert listeners. The left-hand panel shows an example of the manner in which the utterances were visually presented to the listeners. The right-hand panel shows the translation

variants for the words in the lexicon using the five phonological rules described earlier. Pronunciation variants were only generated for the 379 words under investigation, for the other words present in the 186 utterances the canonical transcription was sufficient. The canonical phone transcription is the phone transcription generated with the Text-to-Speech system developed at the University of Nijmegen (Kerkhoff & Rietveld, 1994). The language model (unigram and bigram) was restricted in that it only contained the words present in the utterance which was being recognized.

Feature extraction was done every 10 ms for frames with a width of 16 ms. The first step in feature analysis was an FFT analysis to calculate the spectrum. Next, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz was calculated. The next processing stage was the application of a discrete cosine transformation on the log filterband coefficients. Besides 14 cepstral coefficients ($c_0 - c_{13}$), 14 delta coefficients were also used. Thus, a total of 28 feature coefficients were used.

The acoustic models which we used are monophone hidden Markov models (HMM). The topology of the HMMs is as follows: Each HMM is made up of six states, and consists of three parts. Each of the parts has two identical states, one of which can be skipped (Steinbiss et al., 1993). In total, 40 HMMs were trained. For 33 of the phonemes, one context-independent HMM was used. For the /l/ and the /r/, separate models were trained depending on their position in the syllable, that is, different models were trained for prevocalic and postvocalic position. In addition to these 37 acoustic models, three other models were trained: an HMM for filled pauses, one for nonspeech sounds and a one-state HMM to model silence. Furthermore, the acoustic models which were used for the automatic transcription task were "retrained" models. Retrained acoustic models, in our case, are HMMs which are trained on a training corpus in which pronunciation variation has been transcribed. This is accomplished by performing forced recognition of the training corpus using a lexicon which contains pronunciation variants, thus adding variants to the training corpus at the appropriate places. Subsequently, the resulting corpus is then used to retrain the HMMs. The main reason for using retrained acoustic models is that we expect these

models to be more precise and therefore better suited to the task. For more details on this procedure see Kessens, Wester, and Strik (1999).

Note that we use monophone models rather than diphone or triphone models although in state-of-the-art recognition systems diphone and triphone models have proven to outperform monophone models. This is the case in a recognition task, but not necessarily in forced recognition.

2.1.4 *Evaluation*

Binary scores. On the basis of the judgments made by the listeners and the CSR, scores were assigned to each item. For each of the rules two categories were defined: (1) "rule applied" and (0) "rule not applied." For 88 words four variants were present, as mentioned earlier. For each of these words two binary scores were obtained, that is, for each of the two underlying rules it was determined whether the rule was applied (1) or not (0). For each of the remaining 291 words one binary score was obtained. Thus, 467 binary scores were obtained for each of the listeners and for the CSR.

Agreement. We used Cohen's kappa (Cohen, 1968) to calculate the degree of agreement between listeners and the CSR. The reason we chose to use Cohen's κ instead of for instance percentage agreement is that the distributions of the binary scores may differ for the various phonological processes, and in that case, it is necessary to correct for chance agreement in order to be able to compare the processes to each other. Cohen's κ is a measure which corrects for chance:

$$\kappa = \frac{(P_o - P_c)}{(1 - P_c)} \qquad -1 \le \kappa \le 1$$

where: P_o = observed proportion of agreement; P_c = proportion of agreement on the basis of chance

Table 2 shows the qualifications for κ-values greater than zero, to indicate how the κ-values should be interpreted (taken from Landis & Koch, 1977).

TABLE 2

Qualifications for κ-values >0

k-value	*qualification*
0.00 – 0.20	slight
0.21 – 0.40	fair
0.41 – 0.60	moderate
0.61 – 0.80	substantial
0.81 – 1.00	almost perfect

Reference transcriptions. In the introduction, we mentioned various strategies that can be used to obtain a reference transcription. In this first experiment, we used the majority vote procedure. Two types of reference transcriptions were composed using the majority vote

procedure: 1) reference transcriptions based on eight listeners, and 2) a reference transcription based on all nine listeners.

The reference transcriptions based on eight listeners were used to compare the performance of each individual listener to the performance of the CSR. For each listener, the reference transcription was based on the other eight listeners. By using a reference transcription based on eight listeners, it is possible to compare the CSR and an individual listener to exactly the same reference transcription, thus ensuring a fair and correct comparison. If, instead, one were to use a reference transcription based on all nine listeners, the comparison would not be as fair because, in effect, the listener would be compared to herself/himself due to the fact that the results of that individual listener would be included in the reference transcription.

Consequently, nine sets of reference transcriptions were compiled each with four different degrees of strictness. The different degrees of strictness which we used were A: a majority of at least five out of eight listeners agreeing, B: six out of eight, C: seven out of eight, and finally D: only those cases in which all eight listeners agree. Subsequently, the degree of agreement for an individual listener with the reference transcription was calculated and the same was done for the CSR with the various sets of reference transcriptions.

The reference transcription based on nine listeners was used to analyze the differences between the listeners and the CSR. In this case, it is also possible to use different degrees of strictness. However, for the sake of brevity, we only show the results for a majority of five out of nine listeners agreeing. The reason for choosing five out of nine is that as the reference becomes stricter, the number of items in it reduces, whereas, for this degree of strictness all items (467) are present.

2.2 *Results*

Analysis of the results was done by carrying out four comparisons. First, pairwise agreement was calculated for the various listeners and for the listeners and the CSR. Pairwise agreement gives an indication of how well the results of the listeners compare to each other and to the results of the CSR. However, as we explained in the introduction, pairwise agreement is not the most optimal type of comparison, as the transcriptions of individual transcribers may be incorrect. To circumvent this problem as much as possible, we used the majority vote procedure to obtain reference transcriptions. Thus, we also calculated the degree of agreement between the individual listeners and a reference transcription based on the other eight listeners and between the CSR and the same sets of reference transcriptions. These results give a further indication of how well the listeners and the CSR compare to each other, but we were also curious whether the same pattern exists for the various phonological processes. Therefore, for the third comparison, the data were split up for the separate processes and the degree of agreement between the CSR and the reference transcriptions was calculated for each of the phonological processes. These data showed that there are indeed differences between the various phonological processes. In an attempt to understand the differences, we analyzed the discrepancies between the CSR and the listeners. In this final analysis, the reference transcription based on a majority of five out of nine listeners agreeing was employed.

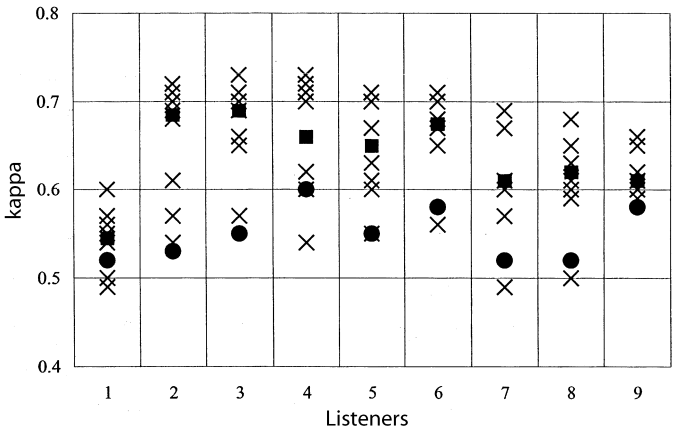


Figure 2
Cohen's κ for the agreement between the CSR and each listener (●), for listener pairs (×) and the median of the listeners (■)

2.2.1
Pairwise agreement between CSR and listeners

For each listener, pairwise agreement was calculated for each pair of listeners and for each CSR-listener pair. In this analysis, no reference transcription was used. Figure 2 shows the results of the pairwise comparisons. For instance, in the first "column" in Figure 2, the crosses (×) indicate the comparison between listener 1 and each of the other listeners, the square (■) shows the median for all listener pairs, and the circle (●) indicates the degree of agreement between the CSR and listener 1.

The results for pairwise agreement in Figure 2 show that there is quite some variation among the different listener pairs. The κ-values vary between 0.49 and 0.73, and the median for all listener pairs is 0.63. The median κ-value for all nine listener-CSR pairs is 0.55. In Figure 2, it can also be seen that the degree of agreement between each of the listeners and the CSR is lower than the median κ-value for the listeners. Statistical tests (Mann-Whitney test, $p < .05$) show that the CSR and listeners 1, 3, and 6 behave significantly different from the other listeners. For both the CSR and listener 1, agreement is significantly lower than for the rest of the listeners whereas for listeners 3 and 6 agreement is significantly higher.

2.2.2
Agreement with reference transcriptions with varying degrees of strictness

In order to further compare the CSR's performance to the listeners', nine sets of reference transcriptions were compiled, each based on eight listeners and with four different degrees of strictness. With an increasingly stricter reference transcription, the differences between listeners are gradually eliminated from the set of judgments under investigation. It is to be expected that if we compare the performance of the CSR with the reference transcriptions of type A, B, C, and D, the degree of agreement between the CSR and the reference transcription will increase when going from A to D. The rationale behind this is that those cases for which a greater number of listeners agree should be easier to judge for the listeners. Therefore, it can be expected that those cases should be easier for the CSR too. In going from A to D the number of cases involved is reduced (see Appendix 1 for details on numbers).

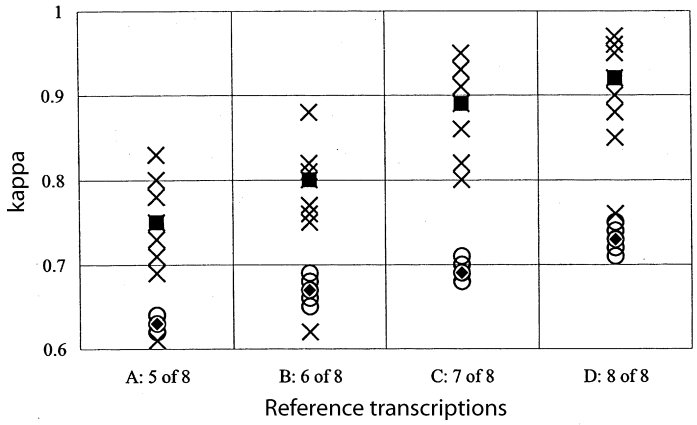


Figure 3

Cohen's κ for CSR (○) and listeners (×) compared to various sets of reference transcriptions based on responses of eight listeners, and median κ for the sets of reference transcriptions for the CSR (◆) and the listeners (■)

Figure 3 shows the κ-values obtained by comparing each of the listener's transcriptions to the relevant set of reference transcriptions (×) and the median for all listeners (■). In addition, the κ-values obtained by comparing the CSR's transcriptions to each of the sets of reference transcriptions (○), and the median for all the CSR's κ-values (◆) are shown. It can be seen that in most cases the degree of agreement between the different sets of reference transcriptions and the listeners is higher than the degree of agreement between the reference transcriptions and the CSR. These differences between the CSR and the listeners are significant. (Wilcoxon signed ranks test, $p < .05$.) However, as we expected, the degree of agreement between the reference transcription and both the listeners and the CSR gradually increases, as the reference transcription becomes stricter.

2.2.3
Agreement with reference transcription for the separate phonological processes

In the previous section, we compared results in which items of the various phonological processes were pooled. However, it is possible that the CSR and the nine listeners perform differently on different phonological processes. Therefore, we also calculated the results for the five phonological processes separately, once again using a majority vote based on eight listeners (see Appendix 2 for the number of items in each set of reference transcriptions). The results are shown in Figure 4. For each process, the degree of agreement between each of the sets of reference transcriptions and the nine listeners (×) and the CSR (○) is shown, first for all of the processes together and then for the individual processes. The median for the nine listeners (■) and the median for the results of the CSR (◆) are also shown. Furthermore, for three of the listeners, the data points have been joined to give an indication of how an individual listener performs on the different processes in relation to the other listeners.

For instance, if we look at the data points for listener A (dotted line) we see that this listener reaches the highest κ-values for all processes except for /n/-deletion in which case the listener is bottom of the group of listeners. The data points for listener B (solid line) fall in the middle of the group of listeners, except for the processes of /r/-deletion and /t/-deletion, where this listener is bottom of the group. The data points for listener C (dashed line) show a poor performance on schwa-insertion and schwa-deletion compared to the

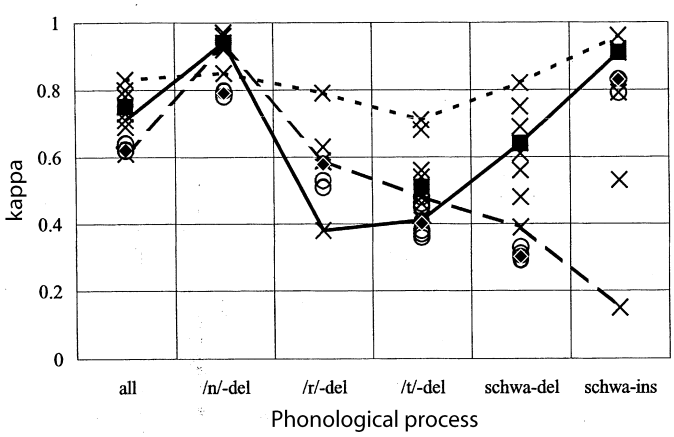


Figure 4
Cohen's κ for the listeners and the CSR compared to the sets of reference transcriptions (5 of 8) for the various phonological processes (○ = CSR, × = listener, ■ = median listeners, ◆ = median CSR, dotted line = listener A, solid line = listener B, and dashed line = listener C)

rest of the listeners, but a more or less average performance on the other processes. These three examples indicate that none of the listeners is consistently better or worse than the others in judging the various phonological processes. Furthermore, on the basis of the medians for the listeners, we can conclude that /n/-deletion and schwa-insertion are the easiest processes to judge, whereas the processes of /r/-deletion, /t/-deletion and schwa-deletion are more difficult processes for listeners to judge. This is also the case for the CSR.

As far as the difference between the CSR and the listeners is concerned, statistical analysis (Wilcoxon signed ranks test, $p < .05$) shows that for the phonological processes of /r/-deletion and schwa-insertion there is no significant difference between the CSR and the listeners. For the other three processes the difference is significant, and this is also the case for all of the phonological processes grouped together. This is also reflected in Figure 4, as there is almost no difference in the median for the CSR and the listeners for /r/-deletion (0.01) and for schwa-insertion (0.08). For /n/-deletion (0.15) and /t/-deletion (0.11), the difference is larger, and comparable to the results found for all rules pooled together (0.12), leaving the main difference in the performance of the listeners and the CSR to be found for schwa-deletion (0.34).

2.2.4
Differences between CSR and listeners

The results in the previous section give rise to the question of why the results are different for various phonological processes and what causes the differences in results between the listeners and the CSR. In this section, we try to answer the question of what causes the discrepancy, by looking more carefully at the differences in transcriptions found for the listeners and the CSR. In these analyses, we used the reference transcription based on a majority of five out of nine listeners agreeing. The reason we use five of nine instead of five of eight is because we wanted to include all of the material used in the experiment in this analysis. Furthermore, instead of using the categorization "rule applied" and "rule not applied" the categories "phone present" and "phone not present" are used to facilitate presentation and interpretation of the data. Each item was categorized according to whether agreement was found between the CSR and the reference transcription or not.

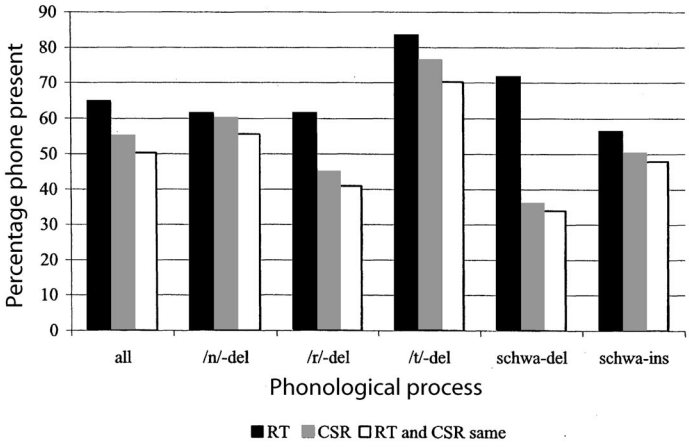


Figure 5

Percentages of phone present for the reference transcription (RT), the CSR, and the CSR and RT together, for the various phonological processes

Figure 5 shows the percentages of phone present according to the reference transcription (RT, dark gray bar) and the CSR (gray bar). It also shows the percentages of phone present for which the RT and CSR agree (white bar). For exact counts and further details, see Appendix 3. It can be seen in Figure 5 that, for all phonological processes pooled, the phones in question are realized in 65% of all cases according to the reference transcription and in 55% of the cases according to the CSR. In fact for every process the same trend can be seen: The RT bar is always higher than the CSR bar. Furthermore, the CSR bar is never much higher than the RT-CSR bar, which indicates that the CSR rarely chooses phone present when the RT chooses phone not present. The differences between the CSR and the listeners are significant for /r/-deletion, for schwa-deletion and for all rules pooled (Wilcoxon signed ranks test, $p < .05$).

An explanation for the differences between the CSR and the listeners may be that they have different durational thresholds for detecting a phone, in the sense that phones with a duration that falls under a certain threshold are less likely to be detected. This sounds plausible if we consider the topology of the HMMs. The HMMs we use have at least three states, thus phones which last less than 30 ms are less likely to be detected. (Feature extraction is done every 10 ms.)

To investigate whether this explanation is correct, we analyzed the data for schwa-deletion and /r/-deletion in terms of the duration of the phones. The speech material was automatically segmented to obtain the durations of the phones. The segmentation was carried out using a transcription that did not contain deletions to ensure that durations could be measured for each phone. Due to the typology of the HMMs durations shorter than 30 ms are also classified as 30 ms As a result, the 30 ms category may contain phones that are shorter in length.

Figures 6 and 7 show the results for schwa-deletion and /r/-deletion, respectively. These figures show that the longer the phone is the less likely that the CSR and the listeners consider it deleted, and the higher the degree of agreement between the CSR and the listeners is. Furthermore, the results for schwa-deletion seem to indicate that the listeners and the CSR do indeed have a different threshold for detecting a phone. Figure 6 shows that the listeners perceive more than 50% of the schwas that are 30 ms or less long, whereas

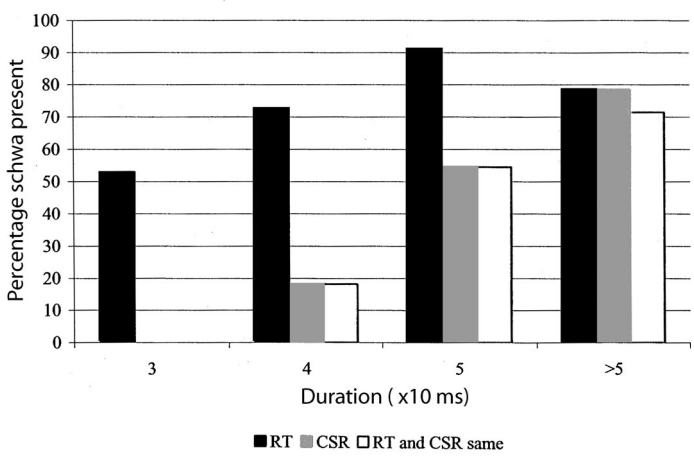


Figure 6

Percentage schwas present, as a function of the duration of the phones, according to the reference transcription (RT), the CSR, and the CSR and RT together

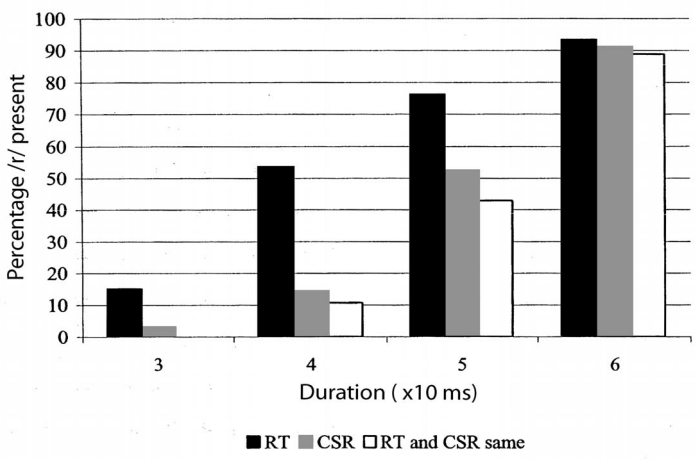


Figure 7

Percentage/r/s present, as a function of the duration of the phones, according to the reference transcription (RT), the CSR, and the CSR and RT together

the CSR does not detect any of them. However, for /r/-deletion this is not quite the case as neither the CSR nor the listeners detect most of the /r/s with a duration of 30 ms or less.

2.3 *Discussion*

The results concerning pairwise agreement between the listeners and the CSR show that the agreement values obtained for the machine differ significantly from the agreement values obtained for the listeners. However, the results of three of the listeners also differ significantly from the rest. Thus, leaving a middle group of six listeners that do not significantly differ from each other. On the basis of these pairwise agreement results, we must conclude that the CSR does not perform the same as the listeners, and what is more that not all of the listeners perform the same either.

A significant difference between the machine's performance and the listeners' performance also appeared when both the CSR transcription and those of the nine listeners were

compared with reference transcriptions of various degrees of strictness. However, the cases that were apparently easier to judge for the listeners, that is, a greater number of them agreed, also presented fewer difficulties for the CSR.

The degrees of agreement observed in this experiment, both between listeners and between listeners and machine, are relatively high. This is all the more so if we consider that the degree of agreement was not calculated over all speech material, as in the Kipp et al. (1997) study, but only for specific cases which are considered to be among the most difficult ones. As a matter of fact, all processes investigated in these experiments are typical connected speech processes that in general have a gradual nature and are therefore difficult to describe in categorical terms (Booij, 1995; Kerswill & Wright, 1990).

In addition, more detailed analyses of the degree of agreement between humans and machine for the various processes revealed that among the phenomena investigated in these experiments there are differences in degree of difficulty. Also in this case the machine's performance turned out to be similar to the listeners', in the sense that the processes that presented more difficulties for the listeners also appeared to be more difficult for the machine. Statistical analyses were carried out for the various phonological processes. The results of these tests are shown in Table 3.

TABLE 3

Results of the statistical analyses for the individual phonological processes from Figure 4 and Figure 5. S = significant; N = not significant difference

Figure	*/n/-deletion*	*/r/-deletion*	*/t/-deletion*	*schwa-deletion*	*schwa-insertion*
4	S	N	S	S	N
5	N	S	N	S	N

Table 3 shows that the comparisons carried out for the individual processes do not present a very clear picture. For schwa-deletion the differences are always significant and for schwa-insertion they are always not significant. For the remaining three processes, the results of the statistical analyses seem to contradict each other. This is maybe less puzzling than it seems if we consider that the comparisons that were made are of a totally different nature. In Figure 4, nine pairs of kappas were compared to each other and in Figure 5, many pairs of "rule applied" and "rule not applied" were compared (the number varies per rule). Still the question remains how we are to interpret these results. The objective was to find out whether the CSR differs significantly from the listeners or not. If we look at the global picture of all rules pooled together then we must conclude that this is indeed the case; the CSR differs significantly from the listeners. However, if we consider the individual processes, we find that the differences for schwa-deletion are significant, for schwa-insertion they are not and that for the other three processes no definite conclusion can be drawn, as it depends on the type of analysis. In other words, only in the case of schwa-deletion are the results of the CSR significantly different from the results of the listeners.

The fact that the degree of agreement between the various listeners and the reference transcriptions turned out to be so variable depending on the process investigated deserves attention, because, in general, the capabilities of transcribers are evaluated in terms of

global measures of performance calculated across all kinds of speech processes, and not as a function of the process under investigation (Shriberg, Kwiatowski, & Hoffman, 1984). However, this experiment has shown that the differences in degree of agreement between the various processes can be substantial.

These results could be related to those presented by Eisen, Tillman, and Draxler (1992) about the variability of interrater and intrarater agreement as a function of the sounds transcribed, although there are some differences in methodology between our experiment and theirs. First, Eisen et al. (1992) did not analyze whether a given segment had been deleted/inserted or not, but whether the same phonetic symbol had been used by different subjects or by the same subject at different times. The degree of agreement in this latter case is directly influenced by the number of possible alternatives, which may be different for the various sounds. In our experiment, on the other hand, this number is constant over all cases. Furthermore, the relative difficulty in determining which particular type of nasal consonant has been realized may be different from the difficulty in determining whether a given nasal consonant is present or not. Second, these authors expressed the degree of agreement using percentage agreement, which, as explained above, does not take chance agreement into account, and therefore makes comparisons rather spurious. In general, however, Eisen et al. (1992) found that consonants were more consistently transcribed than vowels. In our experiment, there is no clear indication that this is the case. Within the class of consonants, Eisen et al. (1992) found that laterals and nasals were more consistently transcribed than fricatives and plosives, which is in line with our findings that higher degrees of agreement were found for /n/-deletion than for /t/-deletion. For liquids no comparison can be made because these were not included in the Eisen et al. (1992) study. As to the vowels, Eisen et al. (1992) found that central vowels were more difficult to transcribe. In our study we cannot make comparisons between different vowel types because only central vowels were involved. In any case, this provides further evidence for the fact that the processes studied in our experiments are among those considered to be more difficult to analyze.

Another important observation to be made on the basis of the results of this experiment is that apparently it is not only the sound in question that counts, be it an /n/ or a schwa, but rather the process being investigated. This is borne out by the fact that the results are so different for schwa-deletion as opposed to schwa-insertion. This point deserves further investigation.

The fourth comparison carried out in Experiment 1 was aimed at obtaining more insight into the differences between the machine's choices and the listeners' choices. These analyses revealed that these differences were systematic and not randomly distributed over presence or absence of the phone in question. Across-the-board the listeners registered more instances of insertion and fewer instances of deletion than the machine did, thus showing a stronger tendency to perceive the presence of a phone than the machine. Although this finding was consistent over the various processes, it was most pronounced for schwa-deletion.

In view of these results, we investigated whether the CSR and the listeners possibly have different durational thresholds in detecting the presence of a phone. This analysis showed that it is clear that duration does certainly play a role, but there is no unambiguous threshold which holds for all phones.

Another possible explanation for these results could be the very nature of the HMMs. These models do not take much account of neighboring sounds. This is certainly true in our case as we used context independent phones, but even when context dependent phone models are used this is still the case. With respect to human perception, on the other hand, we know that the way one sound is perceived very much depends on the identity of the adjacent sounds and the transitions between the sounds. If the presence of a given phone is signaled by cues that are contained in adjacent sounds, the phone in question is perceived as being present by human listeners, but would probably be absent for the machine that does not make use of such cues. A third possible explanation for the discrepancies between the machine response and the listeners' responses lies in the fact that listeners can be influenced by a variety of factors (Cucchiarini, 1993, p. 55), among which spelling and phonotactics are particularly relevant to our study. Since in our experiments the subjects listened to whole utterances, they knew which words the speaker was uttering and this might have induced them to actually "hear" an /r/, a /t/, an /n/ or a schwa when in fact they were not there. In other words, the choice for a nondeletion could indeed be motivated by the fact that the listener knew which phones were supposed to be present rather than by what was actually realized by the speaker. This kind of influence is known to be present even in experienced listeners like those in our experiments. A problem with this argument is that while it can explain the lower percentages of deletion by the humans, it does not explain the higher percentages of insertions. A further complicating factor in our case is that the listeners are linguists and may therefore be influenced by their knowledge and expectations about the processes under investigation. Finally, schwa-insertion happens to be a phenomenon that is more common than schwa-deletion (Kuijpers & Van Donselaar, 1997) which could explain part of the discrepancy found for the two processes.

3 Experiment 2

In Experiment 1, analysis of the separate processes showed that both for listeners and the CSR some processes are more easily agreed on than others. Closer inspection of the differences showed that the CSR systematically tends to choose for deletion (non-insertion) of phones more often than listeners do. This finding was consistent over the various processes and most pronounced for schwa-deletion. Furthermore, we found that the results were quite different for schwa-deletion as opposed to schwa-insertion. To investigate the processes concerning schwa to a further extent, a second experiment was carried out in which we focused on schwa-deletion and schwa-insertion. The first question we would like to see answered pertains to the detectability of schwa: is the difference between listeners and machine truly of a durational nature? In order to try to answer this question, it was necessary to make use of a more detailed transcription in which it was possible for transcribers to indicate durational aspects and other characteristics of schwa more precisely. To achieve this, we used the method of consensus transcriptions to obtain reference transcriptions of the speech material.

The second question is why the processes of schwa-deletion and schwa-insertion lead to such different results. In Experiment 1, the machine achieved almost perfect agreement with listeners on judging the presence of schwa in the case of schwa-insertion, whereas only fair agreement was achieved in the case of schwa-deletion. This difference is quite

large and it is not clear why it exists. Looking at these two processes in more detail could shed light on the matter.

3.1 *Method and Material*

3.1.1 *Phonological variation and selection of speech material*

As was mentioned above, in this second experiment, we concentrated on the phonological processes of schwa-deletion and schwa-insertion. For both processes the material from Experiment 1 was used and both sets were enlarged to include 75 items.

3.1.2 *Experimental procedure*

Listeners. The main difference in the experimental procedure, compared to the previous experiment, is that the consensus transcription method was used instead of the majority vote procedure to obtain a reference transcription. The listeners that participated in this experiment were all Language and Speech Pathology students at the University of Nijmegen. All had attended the same transcription course. The transcriptions used in this experiment were made as a part of the course examination. Six groups of listeners (5 duos and 1 trio, i.e., 13 listeners) were each asked to judge a portion of the 75 schwa-deletion cases and the 75 schwa-insertion cases. The words were presented to the groups in the context of the full utterance. They were instructed to judge each word by reaching consensus of transcription for what was said at the indicated spot in the word (where the conditions for application of the rule were met). The groups were free to transcribe what they heard using a narrow phonetic transcription.

CSR. The CSR was employed in the same fashion as it was in the first experiment; the task was to choose whether a phone was present or not. Because of this, the tasks for the listeners and the machine were not exactly the same. The listeners were not restricted to choosing whether a phone was present or not as the CSR was, but were free to transcribe whatever they heard.

Evaluation. By allowing the listeners to use a narrow phonetic transcription instead of a forced choice, the consensus transcriptions resulted in more categories than the binary categories used previously: "rule applied" and "rule not applied." This is what we anticipated and an advantage in the sense that the transcription is bound to be more precise. However, in order to be compared with the CSR transcriptions, the multivalued transcriptions of the transcribers have to be reduced to dichotomous variables of the kind "rule applied" and "rule not applied." In doing this different options can be taken which lead to different mappings between the listeners' transcriptions and the CSR's and possibly to different results. Below, two different mappings are presented. Furthermore, for the analysis of these data, we once again chose to use the categories "phone present" and "phone not present" to facilitate the comparison of the processes of deletion and insertion.

The transcriptions pertaining to schwa-deletion obtained with the consensus method were: deletion: ø, different realizations of schwa: ə, ɜ̆, ə̲, ə̟, əˠ, and other vowels: ɛ̆, ɜ̈. There were fewer transcriptions pertaining to schwa-insertion, viz.: not present: ø, different realizations of schwa: ə, ɜ̆ and other vowels: ɐ, ɪ. The mappings chosen in this case were based on the idea that duration may be the cause of the difference between man and machine. Thus, for both processes, we used the following two mappings:

I. deletions (ø) are classified as "phone not present" and the rest is classified as "phone present" [ə, ɜ̆, ə̲, ə̟, əˠ, ɛ̆, ɜ̈, ɐ, ɪ]

II. deletions (ø) and short schwas (ɜ̆) are classified as "phone not present" and the rest is classified as "phone present": [ə, ə̲, ə̟, əˠ, ɛ̆, ɜ̈, ɐ, ɪ]

3.2 *Results*

Tables 4 and 5 show the different transcriptions given by the transcribers for schwa-deletion and schwa-insertion, respectively. The first row shows which transcriptions were used, the second row shows the number of times they were used by the transcribers, the third row indicates the number of times the CSR judged the item as phone present and the last row shows the number of times the CSR judged the item as phone not present. These tables show that deletion, schwa and short schwa were used most frequently, thus the choice of the two mappings is justified as the number of times other transcriptions occurred is too small to have any significant impact on further types of possible mappings.

TABLE 4

Reference transcriptions obtained for the process of schwa-deletion, and the classification of these items by the CSR as present or not present

	ø	ə	ɜ̆	ə̲	ə̟	əˠ	ɛ̆	ɜ̈	**total**
RT	18	37	15	1	1	1	1	1	75
phone present	1	21	5	–	1	1	–	1	30
phone not present	17	16	10	1	–	–	1	–	45

TABLE 5

Reference transcriptions obtained for the process of schwa-insertion and the classification of these items by the CSR as present or not present

	ø	ə	ɜ̆	ɪ	ɐ	**total**
RT	32	32	8	2	1	75
phone present	6	28	3	2	–	39
phone not present	26	4	5	–	1	36

Figure 8 shows the percentage of schwas present in the CSR's transcriptions and in the reference transcriptions for the processes of schwa-deletion and schwa-insertion, for both mappings. Comparing the CSR's transcriptions to the reference transcriptions once

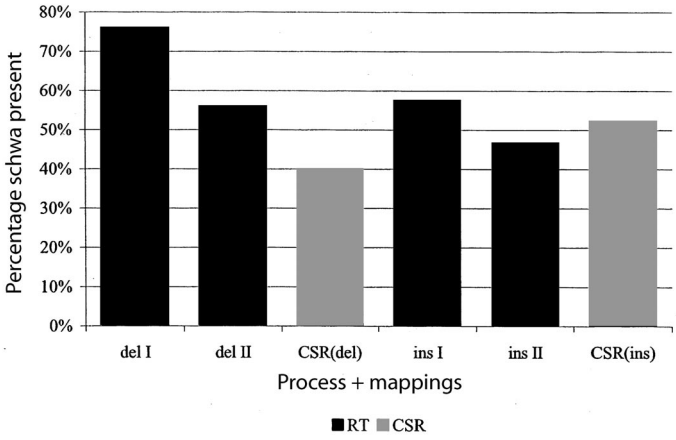


Figure 8

Percentage schwas present for the reference transcription (RT) and for the CSR, for different mappings for the processes of deletion and insertion

again shows that the CSR's threshold for recognizing a schwa is different from the listeners'. In the case of schwa-deletion, this difference becomes smaller when mapping I is replaced by mapping II. For schwa-insertion, replacing mapping I with mapping II leads to a situation where the CSR goes from having a lower percentage of schwa present to having a higher percentage of schwa present than the reference transcription. The difference between the CSR and the reference transcription is significant for schwa-deletion and not significant for schwa-insertion (Wilcoxon, $p < .05$).

Tables 6 and 7 illustrate more precisely what actually occurs. The difference in phone detection between the CSR and the listeners becomes smaller for schwa-deletion (Table 6) if mapping II is used. For this mapping, ə is classified as "phone not present" which causes the degree of agreement between the CSR and the reference transcription to increase. However, it is not the case that all short schwas were classified as "phone not present" by the CSR.

For schwa-insertion (Table 7), the differences in classification by the CSR and by the listeners are not as large. In this case, when the ə is classified as "phone not present" the CSR shows fewer instances of schwa present than the listeners do.

3.3 *Discussion*

The results of this experiment underpin our earlier statement that the CSR and the listeners have different durational thresholds for detecting a phone. A different mapping between the machine and the listeners' results can bring the degree of agreement between the two sets of data closer to each other. It should be noted that the CSR used in this experiment was not optimized for the task, we simply employed the CSR which performed best on a task of pronunciation variation modeling (Kessens, Wester, & Strik, 1999). Although this has not been tested in the present experiment, it seems that changing the machine in such a way that it is able to detect shorter phones more easily should lead to automatic transcriptions that are more similar to those of humans. In other words, in addition to showing how machine and human transcriptions differ from each other, these results also indicate

TABLE 6

Counts of agreement/disagreement CSR and reference transcription (RT) for different mappings of RT categories, for schwa-deletion. **Y**(es) phone present, and **N**(o) phone not present

		RT I			***RT II***		
Mappings		*Y*	*N*	*SUM*	*Y*	*N*	*SUM*
CSR	Y	29	1	30	24	6	30
	N	28	17	45	18	27	45
	SUM	57	18	75	42	33	75

TABLE 7

Counts of agreement/disagreement CSR and reference transcription (RT) for different mappings of RT categories, for schwa-insertion. **Y**(es) phone present, and **N** (o) phone not present

		RT I			***RT II***		
		Y	*N*	*SUM*	*Y*	*N*	*SUM*
CSR	Y	33	6	39	30	9	39
	N	10	26	36	5	31	36
	SUM	43	32	75	35	40	75

how the former could be brought closer to the latter. For instance, the topology of the HMM could be changed by defining fewer states, or by allowing states to be skipped, thus facilitating the recognition of shorter segments.

Although schwa is involved in both cases in this experiment, not much light is shed on the issue of why the processes of insertion and deletion lead to such different results. A possible explanation as far as the listeners are concerned could be the following: For 20 of the schwa-deletion cases, something other than deletion or schwa was transcribed by the listeners compared to nine such cases for schwa-insertion. This indicates that schwa-deletion may be a less straightforward and more variable process. Furthermore, as was mentioned earlier, schwa-deletion is less common than schwa-insertion, which might also influence the judgments of the listeners. So there are two issues playing a role here; the process of deletion might be more gradual and variable than the process of insertion and the listeners may have more difficulties because schwa-deletion is a less frequently occurring process.

Another explanation for the difference is that there is an extra cue for judging the process of schwa-insertion. When schwa-insertion takes place, the /l/ and /r/, which are the left context for schwa-insertion, change from postvocalic to prevocalic position (see Table 8). This change in position within the syllable also entails a change in the phonetic properties of these phones. In general postvocalic /l/s tend to be velarized while postvocalic /r/s tend to be vocalized or to disappear. This is not the case for schwa-deletion, whether or not the schwa is deleted does not influence the type of /l/ or /r/ concerned. These extra cues regarding the specific properties of /l/ and /r/ can be utilized quite easily by listeners, and

TABLE 8

Examples of application of schwa-deletion and schwa-insertion. Syllable markers indicate pre- and postvocalic position of /l/ and /r/

	base form	*rule applied*
schwa-deletion	[la-tə-rə]	[la-trə]
schwa-insertion	[dɛlft]	[dɛ-ləft]

most probably are. They can also be utilized by our CSR because different monophone models were trained for /l/ and /r/ in pre- and post-vocalic position. Thus, whether a schwa is inserted may be easier to judge than whether a schwa is deleted due to these extra cues.

4 General discussion

In this paper, we explored the potential that a technique developed for CSR could have for linguistic research. In particular, we investigated whether and to what extent a tool developed for selecting the pronunciation variant that best matches an input signal could be employed to automatically obtain phonetic transcriptions for the purpose of linguistic research.

To this end, two experiments were carried out in which the performance of a machine in selecting pronunciation variants was compared to that of various listeners who carried out the same task or a similar one. The results of these experiments show that overall the machine's performance is significantly different from the listeners' performance. However, when we consider the individual processes, not all the differences between the machine and the listeners appear to be significant. Furthermore, although there are significant differences between the CSR and the listeners, the differences in performance may well be acceptable depending on what the transcriptions are needed for. Once again it should be kept in mind that the differences that we found between the CSR and the listeners were also in part found between the listeners.

In order to try and understand the differences in degree of agreement between listeners and machine, we carried out further analyses. The important outcome of these analyses is that the differences between the listeners' performance and the machine's did not have a random character, but were of a systematic nature. In particular, the machine was found to have a stronger tendency to choose for absence of a phone than the listeners: the machine signaled more instances of deletion and fewer instances of insertion. Furthermore, in the second experiment, we found that the majority of instances where there was a discrepancy between the CSR's judgments and listeners', it was due to the listeners choosing a short schwa and the CSR choosing a deletion. This underpins the idea that durational effects are playing a role.

In a sense these findings are encouraging because they indicate that the difference between humans and machine is a question of using different thresholds and that by adjusting these thresholds some sort of tuning could be achieved so that the machine's performance becomes more similar to the listeners'. The question is of course whether

this is desirable or not. On the one hand, the answer should be affirmative, because this is also in line with the approach adopted in our research. In order to determine whether the machine's performance is acceptable we compare it with the listeners' performance, which, in the absence of a better alternative, constitutes the point of reference. The corollary of this view is that we should try to bring the machine's performance closer to the listeners' performance. On the other hand, we have pointed out above that human performance does not guarantee hundred percent accuracy. Since we are perfectly aware of the shortcomings of human performance in this respect, we should seriously consider the various cases before unconditionally accepting human performance as the authoritative source.

To summarize, the results of the more detailed analyses of human and machine performance do not immediately suggest that by using an optimization procedure that brings the machine's performance closer to the listeners', better machine transcriptions would be obtained. This brings us back to the point where we started, namely taking human performance as the reference. If it is true that there are systematic differences between human and machine, as appeared from our analyses, then it is not surprising that all agreement measures between listeners were higher than those between listeners and machine. Furthermore, if we have reasons to question the validity of the human responses, at least for some of the cases investigated, it follows that the machine's performance may indeed be better than we have assumed so far.

Going back to the central question in this study, namely whether the techniques that have been developed in CSR to obtain some sort of phonetic transcriptions can be meaningfully used to obtain phonetic transcriptions for linguistic research, we can conclude that the results of our experiments indicate that the automatic tool proposed in this paper can be used effectively to obtain phonetic transcriptions of deletion and insertion processes. It remains to be seen whether these techniques can be extended to other processes.

Another question that arises at this point is how this automatic tool can be used in linguistic studies. It is obvious that it cannot be used to obtain phonetic transcriptions of complete utterances from scratch, but is clearly limited to hypothesis verification, which is probably the most common way of using phonetic transcriptions in various fields of linguistics, like phonetics, phonology, sociolinguistics, and dialectology. In practice, this tool could be used in all research situations in which the phonetic transcriptions have to be made by one person. Given that a CSR does not suffer from tiredness and loss of concentration, it could assist the transcriber who is likely to make mistakes owing to concentration loss. By comparing his/her own transcriptions with those produced by the CSR a transcriber could spot possible errors that are due to absent-mindedness.

Furthermore, this kind of comparison could be useful for other reasons. For instance, a transcriber may be biased by his/her own hypotheses and expectations with obvious consequences for the transcriptions, while the biases which an automatic tool may have can be controlled. Checking the automatic transcriptions may help discover possible biases in the listener's data. In addition, an automatic transcription tool could be employed in those situations in which more than one transcriber is involved; in order to solve possible doubts about what was actually realized. It should be noted that using an automatic transcription tool will be less expensive than having an extra transcriber carry out the same task.

Finally, an important contribution of automatic transcription to linguistics would be that it makes it possible to use existing speech databases for the purpose of linguistic research. The fact that these large amounts of material can be analyzed in a relatively short

time, and with relatively low costs makes automatic transcription even more important (see for instance Cucchiarini & van den Heuvel, 1999). The importance of this aspect for the generalizability of the results cannot be overestimated. And although the CSR is not infallible, the advantages of a very large dataset might very well outweigh the errors introduced by the mistakes the CSR makes.

Received: December 21, 1999; revised manuscript received: October 5, 2000; accepted: December 21, 2000

References

AMOROSA, H., BENDA, U. von, WAGNER, E., & KECK, A. (1985). Transcribing phonetic detail in the speech of unintelligible children: A comparison of procedures. *British Journal of Disorders of Communication*, **20**, 281–287.

BOOIJ, G. (1995). *The phonology of Dutch.* Oxford, U.K.: Clarendon Press.

COHEN, J. A. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

CUCCHIARINI, C. (1993). *Phonetic transcription: A methodological and empirical study*. Ph.D. thesis, University of Nijmegen.

CUCCHIARINI, C., & HEUVEL, H. van den (1999). Postvocalic /r/-deletion in Dutch: More experimental evidence. *Proceedings of the, 14th International Congress of Phonetic Sciences*, San Francisco, **3**, 1673–1676.

CUTLER, A. (1998). The recognition of spoken words with variable representations. *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech: Production and Perception*, Aix-en-Provence, France, 83–92.

DUEZ, D. (1998). The aims of SPoSS. *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech: Production and Perception*, Aix-en-Provence, France, VII–IX.

EISEN, B., TILLMANN, H. G., & DRAXLER, C. (1992). Consistency of judgments in manual labeling of phonetic segments: The distinction between clear and unclear cases. *Proceedings of the International Conference on Spoken Language Processing '92*, Banff, Canada, 871–874.

GREENBERG, S. (1999). Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, **29**(2–4), 159–176.

KEATING, P. (1997). Word-level phonetic variation in large speech corpora. To appear in an issue of *ZAS Working Papers in Linguistics*, Ed. Berndt Pompino-Marschal. Available as <http://www.humnet.ucla.edu/humnet/linguistics/people/keating/berlin1.pdf>.

KERKHOFF, J., & RIETVELD, T. (1994). Prosody in NIROS with FONPARS and ALFEIOS. In P. de Haan & N. Oostdijk (Eds.), *Proceedings of the Department of Language and Speech. University of Nijmegen*, **18**, 107–119.

KERSWILL, P., & WRIGHT, S. (1990). The validity of phonetic transcription: Limitations of a sociolinguistic research tool. *Language Variation and Change*, **2**, 255–275.

KESSENS, J. M., WESTER, M., & STRIK, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication*, **29**(2–4), 193–207.

KUIJPERS, C., & DONSELAAR, W. van (1997). The influence of rhythmic context on schwa epenthesis and schwa deletion in Dutch. *Language and Speech*, **41**(1), 87–108.

KIPP, A., WESENICK, B., & SCHIEL, F. (1997). Pronunciation modeling applied to automatic segmentation of spontaneous speech. *Proceedings of EUROSPEECH '97*, Rhodes, Greece, 1023–1026.

LANDIS, J. R., & KOCH, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

LAVER, J. D. M. (1965). Variability in vowel perception. *Language and Speech*, **8**, 95–121.
MEHTA, G., & CUTLER, A. (1998). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, **31**, 135–156.
OLLER, D. K., & EILERS, R. E. (1975). Phonetic expectation and transcription validity. *Phonetica*, **31**, 288–304.
PYE, C., WILCOX, K. A., & SIREN, K. A. (1988). Refining transcriptions: The significance of transcriber "errors." *Journal of Child Language*, **15**, 17–37.
RISCHEL, J. (1992). Formal linguistics and real speech. *Speech Communication*, **11**, 379–392.
SHRIBERG, L. D., & LOF, L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, **5**, 225–279.
SHRIBERG, L. D., KWIATKOWSKI, J., & HOFFMAN, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research*, **27**, 456–465.
STEINBISS, V., NEY, H., HAEB-UMBACH, R., TRAN, B-H., ESSEN, U., KNESER, R., OERDER, M., MEIER H-G., AUBERT, X., DUGAST, C., & GELLER, D. (1993). The Philips research system for large-vocabulary continuous-speech recognition. *Proceedings of EUROSPEECH '93*, Berlin. Germany, 2125–2128.
STRIK, H., RUSSEL, A., HEUVEL, H. van den, CUCCHIARINI, C., & BOVES, L. (1997). A spoken dialog system for the Dutch public transport information service. *International Journal of Speech Technology*, **2**(2), 119–129.
SWERTS, M., & COLLIER, R. (1992). On the controlled elicitation of spontaneous speech. *Speech Communication*, **11**, 463–468.
TING, A. (1970). Phonetic transcription: A study of transcriber variation. *Report from the Project on Language Concepts and Cognitive Skills Related to the Acquisition of Literacy* (Madison: Wisconsin University).
WESTER, M., KESSENS, J. M., & STRIK, H. (1998). Two automatic approaches for analyzing the frequency of connected speech processes in Dutch. *Proceedings of the International Conference on Spoken Language Processing*, Sydney, **7**, 3351–3356.
WITTING, C. (1962). On the auditory phonetics of connected speech: Errors and attitudes in listening. *Word*, **18**, 221–248.

Appendix 1

Number of items in each reference transcription set per excluded listener

	Set of reference transcriptions								
RT Strictness	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
5 of 8	445	448	449	443	449	454	453	454	448
6 of 8	407	399	395	403	407	399	403	404	398
7 of 8	353	349	340	341	345	338	347	348	354
8 of 8	273	249	251	256	250	250	262	254	258

Appendix 2

Number of items in each reference transcription set per excluded listener for each of the phonological processes. (Strictness: 5 out of 8 listeners agreeing)

	Set of reference transcriptions								
Phonological processes	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
/n/-del	152	151	155	151	153	152	154	153	154
/r/-del	116	120	115	114	117	120	117	121	118
/t/-del	79	80	81	79	80	82	82	80	78
schwa-del	51	50	51	51	51	52	53	52	51
schwa-ins	47	47	47	48	48	48	47	48	47

Appendix 3

Counts (percentages between brackets) of agreement/disagreement CSR and reference transcription (RT) based on a majority of 5 of 9 listeners agreeing, for all items together and split up for each of the processes. Phone present = Y, and phone not present = N

	phonological processes					
	all	***/n/-del***	***/r/-del***	***/t/-del***	***schwa-del***	***schwa-ins***
RT=Y, CSR=Y	235 (50)	86 (55)	52 (41)	59 (70)	18 (34)	23 (48)
RT=N, CSR=N	143 (31)	53 (34)	44 (35)	9 (11)	14 (26)	20 (42)
RT=Y, CSR=N	67 (14)	9 (6)	26 (20)	11 (13)	20 (38)	4 (8)
RT=N, CSR=Y	22 (5)	7 (5)	5 (4)	5 (6)	1 (2)	1 (2)
Total RT=Y	302 (65)	95 (61)	78 (61)	70 (83)	38 (72)	27 (56)
Total CSR=Y	257 (55)	93 (60)	57 (45)	64 (76)	19 (36)	24 (50)
Total items	467 (100)	155 (100)	127 (100)	84 (100)	53 (100)	48 (100)

Publication 3.

M. Wester (2001). Pronunciation modeling for ASR – knowledge-based and data-derived methods. *Submitted to Computer Speech and Language.*

Pronunciation Modeling for ASR - Knowledge-based and Data-derived Methods

MIRJAM WESTER

A^2RT, Department of Language and Speech,
University of Nijmegen, The Netherlands

Abstract

This article focuses on modeling pronunciation variation in two different ways: data-derived and knowledge-based. The knowledge-based approach consists of using phonological rules to generate variants. The data-derived approach consists of performing phone recognition, followed by smoothing using decision trees (D-trees) to alleviate some of the errors in the phone recognition. Using phonological rules led to a small improvement in WER; a data-derived approach in which the phone recognition was smoothed using D-trees prior to lexicon generation led to larger improvements compared to the baseline. The lexicon was employed in two different recognition systems: a hybrid HMM/ANN system and a HMM-based system, to ascertain whether pronunciation variation was truly being modeled. This proved to be the case as no significant differences were found between the results obtained with the two systems. Furthermore, we found that 10% of variants generated by the phonological rules were also found using phone recognition, and this increased to 28% when the phone recognition output was smoothed by using D-trees. This indicates that the D-trees generalize beyond what has been seen in the training material, whereas when the phone recognition approach is employed directly, unseen pronunciations cannot be predicted. In addition, we propose a metric to measure confusability in the lexicon. Using this confusion metric to prune variants results in roughly the same improvement as using the D-tree method.

1 Introduction

It is widely assumed that pronunciation variation is one of the factors which leads to less than optimal performance in automatic speech recognition (ASR) systems. Therefore, in the last few decades, effort has been put into finding solutions to deal with the difficulties linked to pronunciation variation. "Pronunciation variation" as a term could be used to describe most of the variation present in speech. The task of modeling it could consequently be seen as the task of solving the problem of ASR. However, this article has no pretension of going quite that far, seeing as we are not dealing with the full scope of pronunciation variation, but have restricted ourselves to pronunciation variation that becomes apparent in a careful broad phonetic (phonemic) transcription of the speech, in the form of insertions, deletions or substitutions of phones relative to the canonical transcription of the words. This type of pronunciation variation can be said to occur at the segmental level.

Although it is assumed that pronunciation variation, in general, constitutes a problem for ASR, one may wonder if this assumption is correct, and whether modeling pronunciation

variation at the segmental level has anything to offer towards the improvement of ASR performance. In two recent studies (McAllaster et al. 1998; Saraçlar et al. 2000), this question has been addressed. Both come to the conclusion that large improvements are feasible, provided that it is clear exactly which variants occur in the testing data. In McAllaster et al. (1998), the potential significance of accurate pronunciation models was demonstrated on simulated data. If the acoustic observations are matched to the phonemic representations contained in the lexicon, performance can be improved quite dramatically. However, results on real speech were much less spectacular. McAllaster et al. (1998) ascribe this to a mismatch between real speech and the models built from it. In Saraçlar et al. (2000), cheating experiments were conducted by carrying out an unconstrained phone recognition of the test material. Next, an alignment of the phone string with reference word transcriptions was carried out to obtain observed pronunciations. The observed pronunciations were used to augment the lexicon. Rescoring a lattice obtained with an ASR system using the new lexicon showed that a substantial gain in performance is possible if, once again, one can accurately predict word pronunciations.

Thus, it seems that the problem of modeling pronunciation variation lies in accurately predicting the word pronunciations that occur in the test material. In order to achieve this, the pronunciation variants must first be obtained in some way or other. Approaches that have been taken to modeling pronunciation variation can be roughly divided into pronunciation variants derived from a corpus of pronunciation data or from pre-specified phonological rules based on linguistic knowledge (Strik and Cucchiarini 1999). Both have their pros and cons. For instance, the information from linguistic literature is not exhaustive; many processes that occur in real speech are yet to be described. On the other hand, the problem with an approach that employs data to access information is that it is extremely difficult to extract *reliable* information from the data.

Irrespective of how the pronunciations are obtained, choices must be made as to which variants to include in the lexicon, and/or to incorporate at other stages of the recognition process. Simply adding pronunciations en masse is futile, it is all too easy to increase the word error rates (WERs). Predicting which pronunciations will be the correct ones for recognition goes hand in hand with dealing with confusability in the lexicon, which increases when variants are added. Confusability is often introduced by errors in phonemic transcriptions. These phonemic transcriptions are used as the information source from which new variants are derived, consequently incorrect variants may be created. One commonly used procedure to alleviate this is to smooth the phonemic transcriptions - whether provided by linguists (Riley et al. 1999) or phone recognition (Fosler-Lussier 1999) - by using decision trees to limit the observed pronunciation variation. Other approaches (Sloboda and Waibel 1996; Torre et al. 1996) combat confusability by rejecting variants that are highly confusable on the basis of phoneme confusability matrices. In Holter and Svendsen (1999), a maximum likelihood criterion is used to decide which variants to include in the lexicon. In this work, we employ a metric that calculates the confusability in a lexicon, given a set of training data. This metric, which was first introduced in Wester and Fosler-Lussier (2000), is used to compare different lexica with each other and it is also employed to remove confusable variants from lexica.

The first objective of this study is to compare two methods of modeling pronunciation variation which differ at the level of how the pronunciations are obtained. First of all, we look at using phonological rules to obtain pronunciations: the knowledge-based approach. Secondly, we gather information on pronunciations from the data: the data-derived approach to modeling pronunciation variation.

Not only are we interested in obtaining the variants, but we are also interested in incorporating the correct variants in the recognition system, as the studies by McAllaster et al. (1998) and Saraçlar et al. (2000) showed is paramount. Therefore, the second objective is to select those variants produced by the data-derived approach that describe the variance in the data best, but do not lead to errors because of increased confusability within a lexicon. This issue is addressed by calculating the confusability of individual variants in a lexicon on the basis of a forced alignment of the training data using the lexicon for which confusability is to be determined (Wester and Fosler-Lussier 2000). Next, those variants which are earmarked by the confusability metric as highly confusable are discarded from the lexicon, thus creating a lexicon which should contain less confusable variants.

The third objective of this study is to determine whether WER results obtained with a certain lexicon are possibly recognizer dependent. To this end, we compare the effect of one and the same lexicon in two different recognition systems: a hybrid ANN/HMM system and an HMM recognition system. The reason for making a comparison between two different recognition systems is not to find out if one performs better than the other, but to ascertain whether pronunciation variation is truly being modeled. Especially in a data-derived approach there is the potential that a large degree of circularity exists: a certain recognizer is used to carry out a phone recognition, the output of the phone recognition is subsequently used to generate variants, and then the same recognizer is used to test whether incorporating the variants in the recognition process leads to an improvement in WER. The question that arises in this case is whether pronunciation variation is being modeled or if the system is merely being tuned to its own idiosyncrasies. By using the same lexicon in two different recognition systems this can be evaluated.

The merit of the different approaches to modeling pronunciation variation is evaluated by comparing WER results. In addition, we also compare the lexica obtained through the different approaches to analyze how much of the pronunciation variation in a given speech database is modeled by the approaches.

In the following section, the speech material is described. This is followed by a description of the standard set-up of the two recognition systems: the ICSI hybrid ANN/HMM speech recognition system (Bourlard and Morgan 1993) and the Phicos recognition system (Steinbiss et al. 1993). In section 3, the baseline results of the two systems are presented. Next, a description is given of how the various lexica pertaining to pronunciation modeling are created: the knowledge-based approach to generating new pronunciations and the data-derived approach to pronunciation modeling. In Section 5, an extended description of the confusability metric, proposed in Wester and Fosler-Lussier (2000), is given. This is followed by the results of recognition experiments employing the different pronunciation lexica. In section 7, comparisons are made as to which variants overlap in the different lexica. Fi-

nally, we end by discussing the implications of our results and shortly summarizing the most important findings of this research.

2 Material and Recognizers

2.1 Speech Material

In this study, we focus on segmental (phonemic) variation within VIOS (Strik et al. 1997), a Dutch corpus composed of human-machine "dialogues" in the domain of train timetable information, conducted over the telephone. Our training and test material, selected from the VIOS database, consisted of 25,104 utterances (81,090 words) and 6,267 utterances (20,489 words), respectively. This corresponds to 3531 dialogues, with a total duration of 10h48 speech (13h12 silence), consisting of approximately 60% male and 40% female speakers. Recordings with a high level of background noise were excluded.

Figure 1 shows the cumulative frequency of occurrence of the words in the VIOS training material as a function of word frequency rank. This figure has been included to give a better impression of the composition of the VIOS material. Figure 1 shows that 82% of the training material is covered by the 100 most frequently occurring words. In total, 1104 unique words occur in the training material. The 14 most frequently observed words are all one syllable long and cover 48% of the training material. Furthermore, as the VIOS corpus comprises data collected from a train timetable information system 43% of the words in the lexicon concern station names, which corresponds to 16% of the words in the training material.

2.2 CSRs

As was mentioned in the introduction, we are interested in comparing the effect of modeling pronunciation variation using two different recognizers, to find out if the results obtained with one system can be reproduced by another system, or if the results are possibly system dependent. The main difference between the two CSRs is that in the ICSI system acoustic probabilities are estimated by a neural network instead of by mixtures of Gaussians, as is the case in the Phicos system.

The shared characteristics are the choice of phonemes, used to describe the continuous acoustic stream in terms of discrete units, and the language models that were employed. In both systems, 37 phonemes were employed. For the phonemes /l/ and /r/ a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/)[1]. The other 33 phonemes were context-independent. Models for non-speech sounds and silence were also incorporated in the two CSR systems. The systems use word-based unigram and bigram language models.

The lexicon is the same in both systems, in the sense that it contains the orthography of the words and phone transcriptions for the pronunciations. However, it is different in the

[1] SAMPA-notation is used throughout this article. `http://www.phon.ucl.ac.uk/home/sampa/dutch.htm`

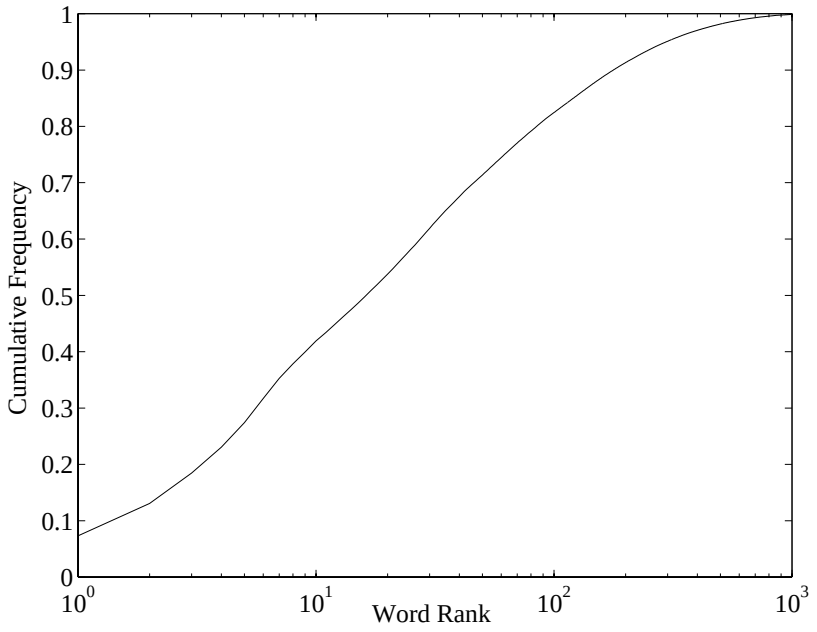


Figure 1: Cumulative frequency of occurrence as a function of word frequency rank for the words in the VIOS training material.

sense that the ICSI lexicon contains prior probabilities for the variants of the words, whereas the Phicos lexicon does not. In the ICSI lexicon the prior probabilities are distributed over all variants for a word and add up to one for each word.

In the Phicos recognition system (Steinbiss et al. 1993), continuous density hidden Markov models (HMMs) with 32 Gaussians per state are used. Each HMM consists of six states, three parts of two identical states, one of which can be skipped. The front-end acoustic processing consists of calculating 14 MFCCs plus their deltas, every 10 ms for 16 ms frames.

The neural network in the ICSI hybrid HMM/ANN speech recognition system (Bourlard and Morgan 1993) was bootstrapped using segmentations of the training material obtained with the Phicos system. These segmentations were obtained by performing a Viterbi alignment using the baseline lexicon (§3.1) and Phicos baseline acoustic models, i.e. no pronunciation variation had been explicitly modeled. For the front-end acoustic processing we use 12^{th}-order PLP features (Hermansky 1990), and energy, which are calculated every 10 ms, for 25 ms frames. The neural net takes acoustic features plus additional context from eight surrounding frames of features at the input, and outputs phoneme posterior probability estimates. The neural network has a hidden layer size of 1000 units and the same network was employed in all experiments.

3 Baseline

In this section, the starting point of our research is outlined. First, the baseline lexicon is described. In Section 3.2, the baseline experiments are described. Baseline here pertains to the condition in which no explicit pronunciation modeling has been carried out. We report on experiments using different feature descriptions for the Phicos system. These experiments were necessary because the standard implementations of the two systems led to significantly different WERs.

3.1 Baseline lexicon

The baseline lexicon comprises 1198 words and contains *one* variant per word. The transcriptions were obtained using the transcription module of a Dutch Text-to-Speech system (Kerkhoff and Rietveld 1994), which looks up the words in two lexica: CELEX (Baayen 1991) and ONOMASTICA, which was used specifically for station names (Quazza and van den Heuvel 2000). For those words for which no transcription was available a grapheme-to-phoneme converter was used, and all transcriptions were manually checked and corrected when necessary. In the ICSI baseline lexicon all prior probabilities are equal to one, as there is only one variant per word. The Phicos lexicon does not contain prior probabilities.

3.2 Baseline results for the ICSI and Phicos recognition systems

Baseline experiments were carried out for the ICSI and Phicos systems. For both systems the "standard" configurations were used. This means that for the Phicos baseline system the feature description consists of 14 MFCCs plus their first derivatives. For the ICSI system the acoustic signal is described using 12 PLP features and energy. Table 1 shows the results of the baseline experiments. The WER for the Phicos system is 12.8% and for the ICSI system it is 10.7%; the difference between these WER results is significant.[2]

In the Phicos standard training procedure, acoustic models are initialized using a linear segmentation. To determine whether using the segmentation produced by the Viterbi alignment may explain part of the difference in WER results, we carried out an experiment in which we substituted the linear segmentation by the same segmentation we used to bootstrap the ICSI system. WERs show that bootstrapping does not have a significant impact in the Phicos system. (For the Phicos linear segmentation the WER is 12.8% and for the Phicos bootstrap segmentation the WER is 12.6%)

To ascertain whether the difference can be explained by the fact that the two systems use different feature descriptions, further experiments were carried out for the Phicos system. First, the same feature description was used as for the ICSI system: 12^{th}-order PLP features and energy. Table 1 shows that this leads to a significant deterioration in WER. The WER result is much higher than the result obtained using 12^{th}-order PLP features and energy in the ICSI system, and the result is also much worse than when MFCCs plus deltas are employed

[2] To establish significance a difference of proportions test was used, with a threshold of .05.

Table 1: Baseline results for ICSI and Phicos systems. Bold indicates that the results differ significantly from the result for the testing condition using 14 MFCCs.

system	**features**	**WER**
ICSI	12 PLPs + e	**10.7**
Phicos	14 MFCCs + Δ	12.8
	12 PLPs + e	**32.3**
	12 PLPs + e + Δ	**11.4**
	12 PLPs + e + Δ + $\Delta\Delta$	**10.4**

in the Phicos system. However, these comparisons are not fair as the amount of context that is taken into consideration in the different systems is not equal. In the original Phicos set-up, context information is incorporated in the acoustic models by using deltas. In the ICSI system, context is dealt with by using the adjacent four frames to the left and right of the frame that is being looked at, as input to the neural net. Therefore, in a subsequent experiment deltas were added, which makes it possible to make a fairer comparison with a Phicos system that uses feature vectors consisting of 14 MFCCs plus deltas. In addition, double deltas were added in order to be able to make the comparison between the ICSI and Phicos systems as fair as possible.

Table 1 shows that 12^{th}-order PLP features plus energy and deltas leads to a significant improvement over 14 MFCCs plus deltas in the Phicos system. The difference between the ICSI and Phicos systems is still significant, with the ICSI system outperforming the Phicos system. However, when double deltas are added to the feature vector the results for the ICSI and Phicos systems are comparable. The result for the Phicos system is slightly better than for the ICSI system, but this difference is not significant.

Thus, it seems the difference in WER between the two systems in their standard configurations can be explained by the different feature descriptions. The conclusion of these experiments is that the improvement in baseline result is both due to using 12^{th}-order PLP features instead of 14 MFCCs, and employing extra context information. The feature descriptions which are used in the rest of the experiments reported on in this paper are 12^{th}-order PLP features and energy for the ICSI system and 12^{th}-order PLP features, with their first and second derivatives, and energy for the Phicos system.

4 Lexica Generation

Using a knowledge-based approach and a data-derived approach to pronunciation modeling, we generated a number of new lexica. In all the newly generated lexica, pronunciation variants were *added* to the baseline lexicon (§3.1). Section 4.1 describes the linguistically motivated approach to modeling pronunciation variation, followed by an explanation of how we derived pronunciations from the data in Section 4.2.

Table 2: Phonological rules and context for application.

Rule	Context for application
/n/-deletion	n → ø/ @ __ #
/r/-deletion	r → ø/ [+vowel] __ [+consonant]
/t/-deletion	t → ø/ [+obstruent] __ [+consonant]
schwa-deletion	@ → ø/ [+obstruent] __ [+liquid][@]
schwa-insertion	ø → @/ [+liquid] __ [-coronal]

4.1 Knowledge-based lexicon

In a knowledge-based approach, the information about pronunciations is derived from knowledge sources, for instance hand-crafted dictionaries or the linguistic literature. In this study, we selected five phonological processes, which are described in the literature, to formulate rules with which pronunciation variants were generated. The rules are context dependent and are applied to the words in the baseline lexicon. The resulting variants are unconditionally added to the lexicon. Table 2 shows the five phonological rules and their application contexts. For a more detailed description of the phonological processes see Kessens et al. (1999).

In the ICSI recognizer, each pronunciation is assigned a prior probability which is usually estimated from the frequency count of the pronunciations seen in the training corpus. However, for the knowledge-based approach we did not base the priors on the training data, but distributed the probability mass evenly over all the pronunciations of a word. This was done in order to be able to make the comparison with the same lexicon used in Phicos as fair as possible (recall Phicos does not contain priors in the lexicon).

4.2 Data-derived lexicon

In a data-derived approach, the information used to develop the lexicon is in some way distilled from the training data. In the following paragraphs, we discuss how we obtained our information about pronunciation variation through phone recognition and subsequently how decision trees (D-trees) are used to smooth the phone recognition output.

4.2.1 Phone recognition

The raw information we used for data-derived generation of lexica was obtained by performing phone recognition of the training material with the ICSI recognizer. In this type of recognition task, the lexicon does not contain words, but a list of 39 phones, and a *phone* bigram grammar is used to provide phonotactic constraints. The output is a sequence of phones; no word boundaries are included. To obtain word boundaries, the phone recognition output is aligned to the reference transcription which does contain word boundaries. The reference transcription is obtained by looking up the transcriptions of the words in the baseline lexicon. A distance measure based on binary phonetic features was employed to align the strings of

phones and insert the word boundaries at the most appropriate places in the string (Fosler-Lussier 1999, pp. 40-41).[3] These alignments are used as the basic information for generating the data-derived lexica.

4.2.2 D-trees

Pronunciation variants obtained from phone transcriptions are at once too many and too few. Thus, one would want to derive some kind of "rules" from the data. The approach we use is based on the decision-tree (D-tree) pronunciation modeling approach developed by Riley and Ljolje (1996) and which has been used by many others in the field (Fosler-Lussier 1999; Riley et al. 1999; Saraçlar et al. 2000; Robinson et al. 2001) for pronunciation modeling of read and spontaneous English.

D-trees are used to predict pronunciations based on the alignment between the reference transcription of the training material and a transcription obtained using phone recognition output. The D-trees are used to smooth the phone recognition output before generating a lexicon. We used the Weka package[4](Witten and Frank 2000) to generate relatively simple D-trees, only taking into account the left and right neighboring phone identity in order to match the type of contexts used in our "phonological rules". According to Riley et al. (1999) most of the modeling gain for the pronunciation trees comes from the immediate +/- 1 phonemic context, lexical stress and syllable boundary location information. Therefore, in a subsequent experiment we also added syllable position (onset, nucleus, coda) as a feature in designing the D-trees. We did not incorporate stress as work by van Kuijk and Boves (1999) showed that information contained in the abstract linguistic feature "lexical stress" deviates too much from realized stress patterns in Dutch data. Therefore, in order to be able to effectively use information pertaining to stress, we would have needed data which had been transcribed at that level.

For each of the 38 phones a D-tree was built. The D-tree model is trying to predict:

$$P(realization \mid canonical, context) \tag{1}$$

by asking questions about the context. Using the distributions in the D-trees, finite state grammars (FSG) were built for the utterances in the training data. During this FSG construction, transitions with a probability lower than 0.1 were disallowed. This results in fewer arcs in the FSG and consequently the possibility of creating spurious pronunciations is diminished. (For instance, not using this pruning step results in a lexicon with 10,754 entries, compared to 5880 entries when a value of 0.1 is used.) Subsequently, the FSG were realigned with the training data, and the resulting "smoothed" phone transcriptions were used to generate a new lexicon.

[3]Instead of using SPE features as in Fosler-Lussier (1999) a categorization based on IPA features was used. The following features were used: voiced, vocalic, consonantal, mid, open, front, central, rounded, diphthong, plosive, fricative, nasal, labial, dental, alveolar, palatal, velar, uvular, glottal, lateral, approximant, and trill.

[4]Weka is a java-based collection of machine learning algorithms for solving real-world data mining problems. `http://www.cs.waikato.ac.nz/ml/weka/index.html`

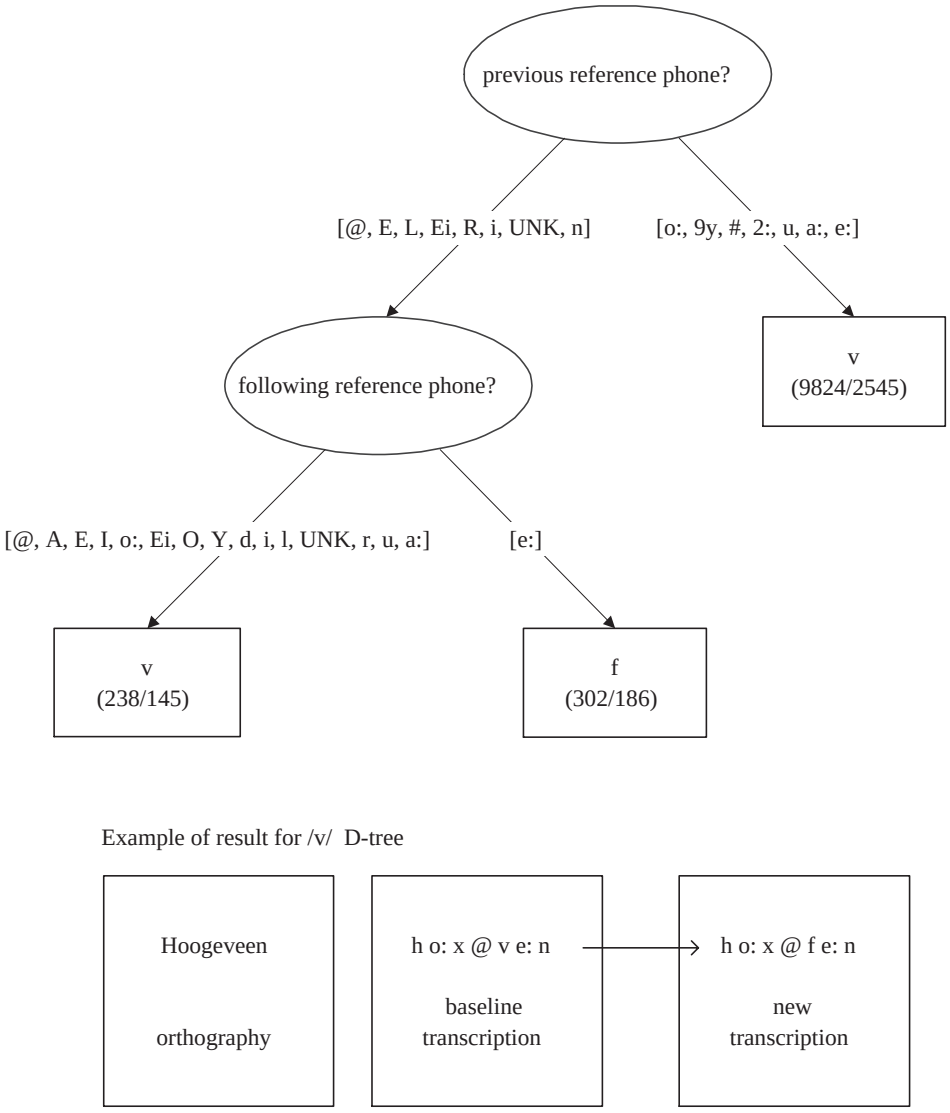


Figure 2: Example of the D-tree generated for the phone /v/, using left and right phone context (UNK = unknown and # = word boundary).

Figure 2 shows an example of a D-tree for the phone /v/. The ovals indicate the questions that are asked, between square brackets the possible answers to the questions are listed. The leaves of the D-tree are depicted by rectangles and contain the outcome of the D-tree. In the example, the outcome is either /v/ or /f/. Two numbers are also shown in the leaves; the first one indicates the number of instances that end up in that leaf and where the phone recognition corresponds to the phoneme given in the leaf. The number following the slash indicates those instances that end up in the leaf, but in which the phone recognition transcription does not correspond to the phoneme given in the leaf. Thus, in this example there are 10,364 instances of /v/ in total, of which 7488 are concordant with the result in the leaf and 2876 that are something other than /v/ or /f/. An example of a variant that could be generated as a result of this D-tree is also shown in Figure 2; /h o: x @ f e: n/ as one of the variants of the station name "Hoogeveen".

4.2.3 Priors in lexicon

Various lexica were generated using the techniques described above. In all cases the prior probabilities for the pronunciations were based on the combination of the phone recognition transcript and pronunciations in the baseline lexicon. The two "lexica" were merged as follows to generate prior probabilities:

$$P_{merged}(pron|word) = \frac{P_{ph.rec.}(pron|word) + P_{baseline}(pron|word)}{2} \quad (2)$$

In the phone recognition lexicon, the probability of a pronunciation is estimated on the basis of the phone recognition transcript ($P_{ph.rec.}$). In the baseline lexicon the probability of the pronunciation of a word is 1 ($P_{baseline}$). When these two lexica are merged, the prior probabilities are re-estimated simply be dividing the priors for the pronunciations of a word by two. When a baseline pronunciation occurs in the phone recognition transcript, it is added to the baseline prior probability and divided by two. Merging in this way ensures that the baseline pronunciations are always present in the new lexicon and that the different lexica contain the same words. If the phone recognition output was taken as is, the result would be out-of-vocabulary words in the testing condition.

5 A measure of confusability

One of the problems that remains at the heart of every approach to modeling pronunciation variation is which variants to include in the lexicon and which to exclude. Some variants lead to improvements and others to deteriorations, and it is difficult to determine which will influence the WER most (Wester et al. 2000). Ideally, what one would want in designing a lexicon is being able to judge beforehand what the optimal set of variants will be for describing the variance in the corpus at the level of the different pronunciations. We took a step in this direction by creating a metric by which we could judge the confusability of individual

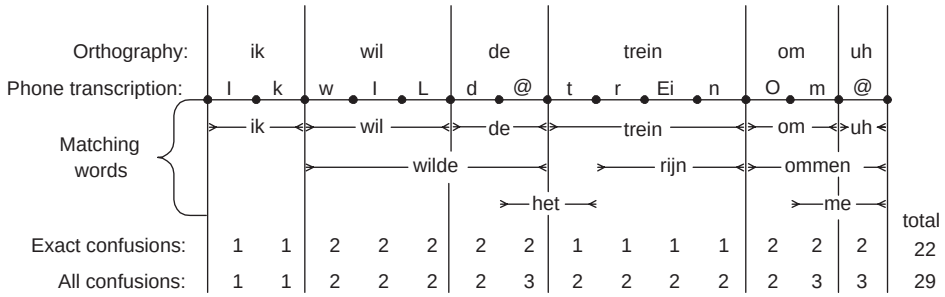


Figure 3: Example of part of the lattice used to compute the average confusion.

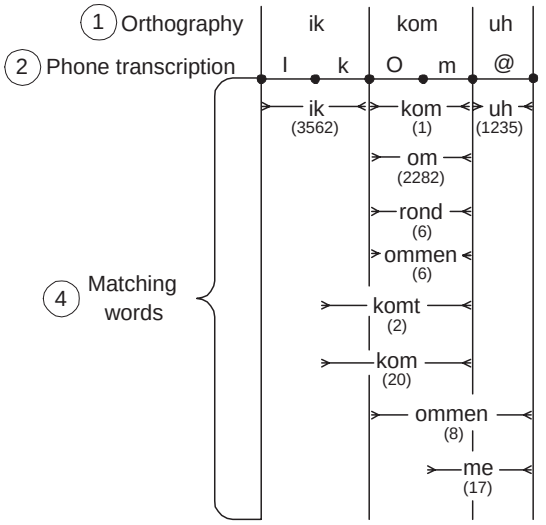
variants, as well as the overall confusability of a lexicon, based on the lexicon containing variants and the training material (Wester and Fosler-Lussier 2000).

The metric works as follows: first a forced alignment of the training data is carried out, using the pronunciations from the lexicon for which the confusability is to be determined. The forced alignment results in a phone transcription of the training material; it should be clear that the phone transcription depends on the variants contained in the lexicon and the acoustic signal. After the phone transcription is obtained, the set of variants that match any substring within the phone alignment is calculated, producing a lattice of possible matching words. For example, in Figure 3, we compute the forced alignment of the word sequence "ik wil de trein om uh" ("I would like to catch the train at uh") resulting in the phonemic string /I k w I L d @ t r Ei n O m @/. We can then find all variants in the lexicon that span any substrings, e.g., the word "wilde" ("wanted" or "wild") corresponding to the phone transcription /w I L d @/.

The confusability metric is calculated by adding up the number of variants that correspond to each phone (as shown in Figure 3 in the row marked "All confusions") divided by the total number of phones. Thus the score for this utterance would be: $\frac{29}{14} = 2.1$, as the total number of phones is 14, and all confusions add up to 29. The average confusability for the lexicon is calculated by summing up the number of words that correspond to each phone in all utterances and dividing by the total number of phones in the training material.

This metric overestimates the number of possible confusions, since it does not take into account that some words would be pruned during decoding because of a dead-end path in the word lattice: for example, the word "het" in Figure 3 does not have an appropriate preceding word in the lattice. The "exact confusion" metric ameliorates this somewhat by not counting words that are stuck in dead-end paths. Since this is an underestimate of the amount of confusion in the lexicon, one can use this as a lower bound of confusability.

In addition to the overall confusability of a lexicon given the training material, we were also interested in obtaining word level confusability scores in order to be able to discard highly confusable variants from the lexicon. The confusability count is defined as the number



3 Lexicon

orthography	*phone transcription*	*count*
ik	I k*	3562
kom	O m	1
kom	k O m*	20
komt	k O m t*	3
komt	k O m	2
me	m @*	17
om	O m*	2282
ommen	O m	6
ommen	O m @*	8
ommen	O m @ n	22
uh	@*	1235
rond	O m	6
rond	r O n t*	62
	*baseline	

Figure 4: Example of part of the lattice used to compute the word confusability scores, and an excerpt from a lexicon containing variants.

of times a variant of a certain word matches the phone transcription of a different word in the training material.

In Figure 4, an example is given to clarify how the word level confusability scores are computed. In this example, the orthography in the training material is "ik kom uh" ① and the phone transcription obtained through forced alignment is /I k O m @/ ②. On the right-hand side in Figure 4, a portion of the lexicon is shown ③. This sample was taken from the lexicon generated using D-trees. The orthography of the words is given, followed by their corresponding phone transcriptions and by a count. This count is the number of times a word is realized in the training material as that specific *variant*. For instance, "kom" is realized as /O m/ in the training material once, and as /k O m/ 20 times.

The words with variants that match the phone transcription ④ are shown in the lattice in Figure 4 with their frequency of occurrence in the training material below between parenthesis. The word level confusability score for "kom" with the corresponding transcription /O m/ is calculated by summing up the counts for all other variants with the same transcription "om(2282)", "rond(6)", and "ommen(6)". Thus the word level confusability count for "kom" /O m/ is 2294. The word level confusability for "om" with the corresponding transcription /O m/ is 13; the sum of "kom(1)", "rond(6)" and "ommen(6)".

In the experiments presented in Section 6.3, a variant of a word is discarded from the lexicon when its confusability count is ≥ 100, unless the variant is the baseline variant. Thus, in this example the variant /O m/ would be discarded for the words "kom", "rond", and

Table 3: Results for the baseline lexicon and lexica generated using the linguistic approach, for the ICSI and Phicos systems.

lexicon	WER ICSI	WER Phicos	variants	vars/ word	conf
Baseline	10.7	10.4	1198	1	1.5
Phon_Rules	10.5	10.3	2066	1.7	1.7
Phon_Rules + LM	10.6	10.2	2066	1.7	1.7
Phon_Rules + LM + PM	10.7	10.1	2066	1.7	1.7

"ommen".

6 Results

This section describes the results that were obtained using the various approaches to modeling pronunciation variation. In an attempt to be as clear as possible, the names of the lexica are indicated in the text and tables in *italics*. The tables show the word error rate (WER) results, the number of entries in the lexica (variants), the average number of variants per word (vars/word) and the confusability of the lexicon (conf), i.e. the average phone level confusion over all words in the training material. Once again, results that differ significantly from the baseline result are indicated in bold. To establish significance a difference of proportions test was used, with a threshold of .05.

6.1 Knowledge-based approach

Phonological rules were used to generate variants, all of which were added to the baseline lexicon to create a new lexicon (*Phon_Rules*). In Kessens et al. (1999), we found that modeling pronunciation variation at all three levels in the recognizer, i.e. the lexicon, the language model and the phone models, led to the largest decrease in error rates within the Phicos recognition system using 14 MFCCs plus deltas. We repeated these experiments for the Phicos system using 12^{th}-order PLP features, with their first and second derivatives, and energy. To discover whether including pronunciation variation at all three levels is also beneficial to the performance of the ICSI system, we incorporated pronunciation variation in the language model by adding probabilities for the pronunciation variants instead of for the words (*Phon_Rules + LM*) and retrained the neural networks on the basis of a new alignment containing the pronunciation variants of the five phonological rules (*Phon_Rules + LM + PM*).

Recall that priors for the variants in the ICSI lexicon are all equal in these experiments to make the comparison with the Phicos system as fair as possible (§ 4.1). Just to make sure the ICSI system is not being penalized by this choice, we also measured the effect of estimating the priors on the training data. We found that using priors estimated on the basis of the training material led to the same WER as using a uniform distribution.

Table 4: Results for lexica generated using a data-derived approach, for the ICSI and Phicos systems.

lexicon	**WER ICSI**	**WER Phicos**	**variants**	**vars/ word**	**conf**
Baseline	10.7	10.4	1198	1	1.5
Phone_Rec	10.9	–	20347	17.7	65.9
D-tree	**9.9**	–	5880	4.9	9.3
D-tree_Syl (no priors)	**17.0**	**17.0**	5912	4.9	9.0
D-tree_Syl + priors/LM	**9.9**	10.0	5912	4.9	9.0
D-tree_Syl + LM + PM	–	10.3	5912	4.9	9.0

Table 3 shows the WER results for the ICSI and Phicos systems when five phonological rules are employed in the recognition system. These results show that modeling pronunciation variation using the five phonological rules has no effect on WERs in the ICSI system, whereas when linguistically motivated pronunciation variation is modeled at all three levels in the Phicos system an improvement is found at each step; however the final result is not significantly better than the baseline result. On the basis of these results, we decided not to add variants to the language model and phone models for the ICSI system in further experiments, whereas we did for the Phicos system.

6.2 Data-derived approach

In the following stage, lexica were created using the data-derived approach (§4.2). First of all, a lexicon was generated on the basis of the "raw" phone recognition output (*Phone_Rec*). Next, a lexicon was generated using D-trees that were created using the phone recognition transcripts and a context consisting of left and right neighboring phones (*D-tree*); and finally a lexicon was created using D-trees which incorporated syllable information in addition to left and right neighboring phones (*D-tree_Syl*).

The *D-tree_Syl* lexicon was used to determine whether a data-derived lexicon generated with one system would lead to similar results when tested in a different system. To this end, the *D-tree_Syl* lexicon was employed in the Phicos system. To ascertain the effect of priors in the ICSI lexica, an experiment was carried out in which the priors in the lexicon were ignored during decoding (*D-tree_Syl (no priors)*). This situation is comparable to the Phicos testing condition in which variants are added to the lexicon only. Next, for the Phicos system pronunciation variants were incorporated in the language model (*D-tree_Syl + LM*), which is comparable to (*D-tree_Syl + priors*) for ICSI. Finally, the phone models were retrained (*D-tree_Syl + LM + PM*) for the Phicos system, as previous experiments with Phicos have shown that the best way of incorporating pronunciation variation is to do it at all three levels. For the ICSI system, this last testing condition was not carried out.

Table 4 shows the WERs for the ICSI and Phicos systems using the different data-derived lexica. Adding all the variants from the raw phone recognition leads to a deterioration in

performance. The deterioration is not as large as one might expect, but it should be kept in mind that the lexicon does not only contain variants from the phone recognition, because, like all other lexica, it was merged with the baseline lexicon and the priors for the baseline variants are higher than the priors for other variants. In any case, the decoding time does increase substantially, which is in line with expectations.

The results in Table 4 further show that modeling pronunciation variation using D-trees leads to a significant improvement in the ICSI system. A relative improvement of 7.5% compared to the baseline result is found. Including syllable information in the D-trees in addition to left and right neighboring phone identity does not further improve the performance.

Simply employing the *D-tree_Syl* lexicon in the Phicos system leads to a significant deterioration in WER compared to the baseline result. Ignoring the priors in the ICSI lexicon leads to a deterioration of the same magnitude. When the variants are added to the language models the performance of the Phicos system improves dramatically, although the improvement is not significant compared to the baseline result. Incorporating pronunciation variation in the recognition process by retraining the phone models leads to a slight degradation compared to only incorporating it in the language models. This is a slightly surprising result as in previous experiments retraining has always led to improvements in WER.

Inspection of the lexical confusability scores in Table 3 and 4 shows that the highest degree of confusability is clearly found in the phone recognition lexica; this is followed by the D-trees lexica, and the least amount of confusability is contained in the phonological rule lexica. However, there is no straightforward relationship between the confusability score and the WER performance. Consequently, it is not clear how the confusability score could be used to predict which lexicon is "better". In addition, there is no relationship between the number of entries in the lexicon (or the number of variants per word) and the WER. However, decoding time increases dramatically with a higher number of entries in the lexicon, which is an extra reason to sparingly add variants to the lexicon. In the following section, we employ the confusability metric to discard confusable variants instead of only measuring the confusability in a lexicon.

6.3 Confusability measure for pruning

The confusability metric was used to prune variants with a confusability count of 100 or higher. For the phone recognition lexicon we also applied a threshold of 0; removing all confusable variants bar the baseline variants. In all cases, the baseline pronunciations were not removed from the lexica. The pruning was applied to the lexica: *Phon_Rules*, *D-trees_Syl*, and *Phone_Rec*. Table 5, column 2 shows the original WERs for the ICSI system prior to pruning with the confusability metric. The remaining columns show results for lexica after pruning had been carried out.

For the *Phon_Rules* lexicon and the *D-tree_Syl* lexicon, pruning the most confusable variants has no effect on the WERs compared to the same testing condition without using the confusability metric to prune variants. This is in contrast to the results found for the "raw" phone recognition lexicon (*Phon_Rec Conf*), where using the confusability metric to prune

Table 5: Results of using confusability metric to remove variants from lexica for the ICSI system.

	without pruning	with pruning			
lexicon	**WER ICSI**	**WER ICSI**	**variants**	**vars/ word**	**conf**
Phon_Rules Conf $\geq$ *100*	10.5	10.5	2054	1.7	1.6
D-tree_Syl Conf $\geq$ *100*	**9.9**	**10.0**	5474	4.6	2.1
Phone_Rec Conf $\geq$ *100*	10.9	**10.1**	15424	12.9	3.2
Phone_Rec Conf $\geq$ *0*	10.9	**10.1**	9222	7.7	1.7

the most confusable variants leads to a significant improvement.

The difference in number of variants present in the phone recognition lexica also deserves some attention. Even when the confusability count for confusable words is set to 0, the *Phone_Rec* lexicon contains almost twice as many variants as the *D_tree* lexicon. This is due to the fact that many of the variants that are generated on the basis of phone recognition are so different from pronunciations chosen during forced alignment that they do not form a match with any of the forced alignment transcriptions. Some other way of pruning these "strange" pronunciations should be employed, as they do not seem to affect the WERs, but they do increase decoding times. It may seem strange that the confusability score for *Phone_Rec Conf* $\geq$ *0* is not 1.5 as it is for the *Baseline* lexicon, but this is due to the fact that after all the confusable variants have been removed, a forced alignment of the training data is carried out again using the new lexicon. As the set of variants is different, the alignments also turns out differently and consequently other variants may be confused with each other.

7 Analysis of Lexica

An analysis was carried out to determine how much overlap there is between lexica generated using the phonological rule method for generating variants and the data-derived approaches to generating variants. The *Phon_Rules* lexicon was used as the starting point for the comparison of the different lexica. This lexicon was chosen because the variants generated by the five phonological rules are valid variants, from a linguistic point of view. From an ASR point of view, the validity of the variants depends on whether the variants actually occur in the data. Therefore, we made comparisons using all variants generated by the phonological rules (Table 6), and only those variants that actually occur in the training material (Table 7).

For each of the phonological rules (see Table 2) lists of variants were made. The extra category "combination" in Table 6 refers to the variants that are the result of more than one rule applying to a word. None of the variants were included in more than one list and baseline variants were not included. The overlap between the lexica was calculated by enumerating the variants (#vars) that occur in both the *Phon_Rules* and *Phone_Rec* lexicon, as well as in the *Phon_Rules* and the *D-tree* lexicon. The percentages indicate the proportion of variants

Table 6: Overlap between variants generated using five phonological rules and variants obtained using data-derived methods.

	lexicon				
	Phon_Rules	*Phone_Rec*		*D-tree*	
rules	#vars	#vars	%	#vars	%
/n/-deletion	283	35	12	104	37
/r/-deletion	240	33	14	80	33
/t/-deletion	63	9	14	23	37
schwa-deletion	19	1	5	4	21
schwa-insertion	65	1	0	2	3
combination	200	10	5	29	15
total	868	89	10	242	28

in the *Phon_Rules* lexicon that is covered by the other lexica.

From the results shown in Table 6, we can infer that the D-trees are learning phonological rules. The *Phon_Rules* column shows that in total 868 variants are generated using the five phonological rules. The phone recognition lexicon, which is based on the raw phone recognition contains only 89 of those variants, which corresponds to 10% of the variants generated by the five phonological rules. The *D_tree* lexicon contains 242 of the 868 variants, which corresponds to 28% of the variants generated by the phonological rules. Thus, 153 new variants are generated by using D-trees to smooth the phone recognition. This is a clear advantage of the D-tree method over simply using the raw phone recognition output to generate variants (although this is much faster, simpler and straightforward). The D-trees manage to generalize beyond what has been seen in the training material, whereas when the phone recognition approach is employed unseen pronunciations cannot be predicted. Another advantage is that the number of variants that is generated by the D-trees is merely a third of the variants present in the *Phone_Rec* lexicon.

In Table 7, the same type of data is presented as in the previous table, with the difference that only those variants that actually occur in the training material are presented. A forced alignment of the training material was carried out using the *Phon_Rules* lexicon to find out which variants actually occur. Table 7 shows that 56% (490/868) of the variants generated by the phonological rules actually occur in the training material. The results further show that almost all of the variants that were generated using D-trees in Table 6 actually occur in the training material when the *Phon_Rules* lexicon is used to carry out a forced alignment. (226 of 242 variants). Thus, the coverage of phonological variants in the D-trees lexicon increases to 46%. For the *Phone_Rec* lexicon the coverage does not increase quite as dramatically, but in this case also almost all of the variants that were found in phone recognition also actually occur in forced alignment using the *Phon_Rules* lexicon.

Table 7: Overlap between variants generated using five phonological rules which truly occur in the training material and variants generated using phone recognition or variants generated by the D-trees.

	lexicon				
	Phon_Rules	*Phone_Rec*		*D-tree*	
rules	#vars	#vars	%	#vars	%
/n/-deletion	195	34	17	100	51
/r/-deletion	141	30	21	77	55
/t/-deletion	37	9	24	20	54
schwa-deletion	13	1	8	4	31
schwa-insertion	36	1	3	2	6
combination	68	10	15	23	34
total	490	85	17	226	46

8 Discussion

In this paper, we reported on two different approaches to dealing with pronunciation variation; a knowledge-based and data-derived approach. The first issue we set out to address was to compare these two approaches to modeling pronunciation variation. The approaches differ in the way that information on pronunciation variation is obtained. The knowledge-based approach consists of generating variants by using phonological rules for Dutch. The data-derived approach consists of performing phone recognition to obtain information on the pronunciation variation in the data, followed by smoothing with D-trees to alleviate some of the *unreliable* data introduced by shortcomings of the recognition system. Both approaches lead to improvements, but of differing magnitudes. The only statistically significant improvement we found, compared to the baseline result, was when we modeled pronunciation variation using a data-derived approach in the ICSI system. However, although the other results do not show a significant improvement over the baseline performance, they also do not differ significantly from the data-derived ICSI result.

Improvements due to modeling pronunciation variation using phonological rules are reported in quite a number of studies (Cohen 1989; Flach 1995; Lamel and Adda 1996; Safra et al. 1998; Wiseman and Downey 1998; Ferreiros and Pardo 1999) for different types of speech, different languages, and employing different CSR systems. Unfortunately, relating the findings in those studies to each other and to the results found in this work is exceedingly difficult because there are factors that may have influenced the findings, but which have not been described in the studies, or which have not been investigated individually. Furthermore, as was stated in Strik and Cucchiarini (1999): "It is wrong to take the change in WER as the only criterion for evaluation, because this change is dependent on at least three different factors: (1) the corpora, (2) the ASR system and (3) the baseline system. This means that improvements in WER can be compared with each other only if in the methods under study these three elements were identical or at least similar." As there is not much else but WERs

to go by it should be clear it is extremely difficult to compare the different studies with each other.

In Kessens et al. (1999) and this study, the exact same training and test data, and CSR were used which makes a comparison possible. In contrast to the results in Kessens et al. (1999), a significant improvement using the knowledge-based approach in Phicos was not found in this study. The difference between the experiments carried out using Phicos is the acoustic features that were employed. In this study, the starting point WER is significantly lower than in Kessens et al. (1999). Our results show that even though the trends are the same, pronunciation modeling through phonological rules has less effect when the starting-point WER is lower. In this case, it seems that the mistakes that were previously solved by modeling pronunciation variation are now being taken care of by improved acoustic modeling. This type of effect is also found in Ma et al. (1998) and Holter and Svendsen (1999). However, there are examples in the literature that this does not necessarily need to be the case. For instance, Riley et al. (1999) reports that reductions in WER due to modeling pronunciation variation persist after the baseline systems are improved by coarticulation sensitive acoustic modeling and improved language modeling.

One of the disadvantages of using a knowledge-based approach, i.e. not all of the variation that occurs in spontaneous speech has been described, is in part alleviated by using a data-derived approach. The challenge that is introduced when a data-derived approach is taken, is that the information which is used to generate variants is not always reliable. Results pertaining to the data-derived approach showed that simply adding all the variants from the raw phone recognition leads to a deterioration in performance. However, when subsequently D-trees were used to smooth the phone recognition, significant improvements in the ICSI system were found. A relative improvement of 7.5% was found compared to the baseline result. This is similar to findings reported for English (e.g. (Fosler-Lussier 1999; Riley et al. 1999; Saraçlar et al. 2000; Robinson et al. 2001)) in the sense that improvements are found when D-trees are used to model pronunciation variation.

One of the other questions we were interested in answering: “Is pronunciation variation indeed being modeled, or are idiosyncrasies of the system simply being modeled?” can be answered by considering the following. First of all, the similar results obtained using two quite different recognition systems indicates that pronunciation variation is indeed being modeled. Although the overall improvements found for the hybrid ANN/HMM system were larger than for the HMM system when using a data-derived approach in which the ANN/HMM system was used to generate the variants and subsequently was used to measure the difference in WERs, the differences between a hybrid ANN/HMM and a standard HMM system were not significant. Secondly, analysis of the lexica showed that the D-trees are learning phonological rules. We found that 10% of variants generated by the phonological rules were also found using phone recognition, and this increased to 28% when the phone recognition output was smoothed by using D-trees. Apparently phonological rule variants are created which were not present in the output of the raw phone recognition. This is a clear advantage of using D-trees over simply using phone recognition output, because the D-trees are capable of generalizing beyond what has been seen in the training material, whereas when the phone

recognition approach is employed unseen pronunciations cannot be predicted. Furthermore, it is an indication that pronunciation variation is indeed being modeled.

Confusability is intuitively an extremely important point to address in pronunciation modeling. The confusability metric which we introduced is useful as a method for pruning variants. The results show that simply pruning highly confusable variants from the phone recognition lexicon leads to a significant improvement compared to the baseline. In other words, the confusability metric is a very simple and easy way of obtaining a result which is comparable to the result obtained using methods such as phonological rules or D-trees. However, we also intended to use the confusability metric to assign a score to a lexicon which could then be used to predict how well a lexicon would perform. The results in Table 5 quite conclusively demonstrate that the confusability score is not suited for this purpose as different confusability scores lead to roughly the same WER scores.

Many studies (e.g. Cohen (1989, Yang and Martens (2000, Ma et al. (1998)) have found that probabilities of the variants (or probabilities of rules) play an important role in whether an approach to modeling pronunciation variation is successful or not. In this study, this was once again shown by comparing results between Phicos and the ICSI system in §6.2. Not including priors in the ICSI system and not incorporating variants in the language model for Phicos showed significant deteriorations, whereas including probabilities showed significant improvements over the baseline. Yet if we are to relate this to the findings of McAllaster et al. (1998) and Saraçlar et al. (2000): if one can accurately predict word pronunciations in a certain test utterance the performance should improve substantially, we must conclude that estimating the priors for a whole lexicon is not optimal. The point is that a good estimation of priors is probably a conditional probability with speaker, speaking mode, speaking rate, subject, etc. as conditionals. Some of these factors can be dealt with in a two-pass scheme by rescoring n-best lists as the pronunciation models in Fosler-Lussier (1999) showed; however, the gains found in this study remain small as it is extremely difficult to accurately estimate the conditionals.

9 Conclusions

A knowledge-based approach for modeling pronunciation variation in Dutch using five phonological rules leads to small improvements in recognition performance. Using a data-derived approach can lead to significant improvements when the phone recognition output is either smoothed by D-trees or pruned using the confusability metric. Using the confusion metric to prune variants results in roughly the same improvement as using the D-tree method. Finally, it is encouraging that using two different recognition systems lead to roughly the same results, as this indicates that pronunciation variation is indeed being modeled and not merely idiosyncrasies of a certain system.

In summary, pronunciation modeling leads to improvements in WER, but not as large as had been hoped. Obtaining accurate predictions of the specific variants that occur in the testing material remains a challenging issue which has not yet been solved.

10 Acknowledgments

I extend my appreciation and gratitude to Eric Fosler-Lussier for his help in getting me acquainted with the ICSI recognizer, and for developing the confusability metric. I would also like to thank Johan de Veth, for his help with the PLP experiments for Phicos. Grateful appreciation is extended to members of A^2RT who gave useful comments on previous versions of this paper, especially, Helmer Strik, Loe Boves, Judith Kessens and Febe de Wet. Furthermore, this work would not have been possible without the hospitality of ICSI — where I was given the opportunity to visit for a number of months — and by a grant from the University of Nijmegen (Frye stipend) and a NWO travel scholarship.

References

Baayen, H. (1991). De CELEX lexicale databank. *Forum der Letteren 32*(3), 221–231.

Bourlard, H. and N. Morgan (1993). *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers.

Cohen, M. (1989). *Phonological Structures for Speech Recognition*. Ph. D. thesis, University of California, Berkeley, CA.

Ferreiros, J. and J. Pardo (1999). Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations. *Speech Communication 29*, 65–76.

Flach, G. (1995). Modelling pronunciation variability for special domains. In *Proc. of EUROSPEECH '95*, Madrid, pp. 1743–1746.

Fosler-Lussier, E. (1999). *Dynamic Pronunciation Models for Automatic Speech Recognition*. Ph. D. thesis, University of California, Berkeley, CA.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) analysis of speech. *Journal of the Acoustic Society of America 87(4)*, 1738–1752.

Holter, T. and T. Svendsen (1999). Maximum likelihood modeling of pronunciation variation. *Speech Communication 29*, 177 –191.

Kerkhoff, J. and T. Rietveld (1994). Prosody in NIROS with FONPARS and ALFEIOS. In P. de Haan and N. Oostdijk (Eds.), *Proc. of the Dept. of Language and Speech, University of Nijmegen*, Volume 18, pp. 107–119.

Kessens, J., M. Wester, and H. Strik (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication 29*, 193–207.

Lamel, L. and G. Adda (1996). On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *Proc. of ICSLP '96*, Philadelphia, PA., pp. 6–9.

Ma, K., G. Zavaliagkos, and R. Iyer (1998). Pronunciation modeling for large vocabulary conversational speech recognition. In *Proc. of ICSLP '98*, Sydney, pp. 2455–2458.

McAllaster, D., L. Gillick, F. Scattone, and M. Newman (1998). Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In *Proc. of ICSLP '98*, Sydney, pp. 1847–1850.

Quazza, S. and H. van den Heuvel (2000). The use of lexicons in text-to-speech-systems. In F. van Eynde and D. Gibbon (Eds.), *Lexicon Development for Speech and Language Processing*, Chapter 7, pp. 207–233. Kluwer Academic Publishers.

Riley, M., W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraçlar, C. Wooters, and G. Zavaliagkos (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication 29*, 209–224.

Riley, M. and A. Ljolje (1996). Automatic generation of detailed pronunciation lexicons. In C.-H. Lee, F. Soong, and K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition: Advanced Topics*, Chapter 12, pp. 285–302. Kluwer Academic Publishers.

Robinson, A., G. Cook, D. Ellis, E. Fosler-Lussier, S. Renals, and D. Williams (2001). Connectionist speech recognition of Broadcast News. *To appear in Speech Communication*.

Safra, S., G. Lehtinen, and K. Huber (1998). Modeling pronunciation variations and coarticulation with finite-state transducers in CSR. In *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, pp. 125–130.

Saraçlar, M., H. Nock, and S. Khudanpur (2000). Pronunciation modeling by sharing Gaussian densities across phonetic models. *Computer Speech and Language 14*, 137–160.

Sloboda, T. and A. Waibel (1996). Dictionary learning for spontaneous speech recognition. In *Proc. of ICSLP '96*, Philadelphia, PA., pp. 2328–2331.

Steinbiss, V., H. Ney, R. Haeb-Umbach, B.-H. Tran, U. Essen, R. Kneser, M. Oerder, H.-G. Meier, X. Aubert, C. Dugast, and D. Geller (1993). The Philips research system for large-vocabulary continuous-speech recognition. In *Proc. of EUROSPEECH '93*, Berlin, pp. 2125–2128.

Strik, H. and C. Cucchiarini (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication 29*, 225–246.

Strik, H., A. Russel, H. van den Heuvel, C. Cucchiarini, and L. Boves (1997). A spoken dialogue system for the Dutch public transport information service. *International Journal of Speech Technology 2*(2), 119–129.

Torre, D., L. Villarrubia, L. Hernández, and J. Elvira (1996). Automatic alternative transcription generation and vocabulary selection for flexible word recognizers. In *Proc. of ICASSP '96*, Munich, pp. 1463–1466.

van Kuijk, D. and L. Boves (1999). Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication 27*, 95–111.

Wester, M. and E. Fosler-Lussier (2000). A comparison of data-derived and knowledge-based modeling of pronunciation variation. In *Proc. of ICSLP '00*, Volume I, Beijing, pp. 270–273.

Wester, M., J. Kessens, and H. Strik (2000). Pronunciation variation in ASR: Which variation to model? In *Proc. of ICSLP '00*, Volume IV, Beijing, pp. 488–491.

Wiseman, R. and S. Downey (1998). Dymanic and static improvements to lexical baseforms. In *Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, pp. 157–162.

Witten, I. and E. Frank (2000). *Data Mining, practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers.

Yang, Q. and J.-P. Martens (2000). On the importance of exception and cross-word rules for the data-driven creation of lexica for ASR. In *Proc. of the 11^{th} ProRisc Workshop*, Veldhoven, The Netherlands, pp. 589–593.

Publication

4.

M. Wester, S. Greenberg and S. Chang (2001). A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, Aalborg, pp. 1729-1732.

A Dutch Treatment of an Elitist Approach to Articulatory-Acoustic Feature Classification

MIRJAM WESTER, STEVEN GREENBERG AND SHUANGYU CHANG

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
{mwester,steveng,shawnc}@icsi.berkeley.edu

Abstract

A novel approach to articulatory-acoustic feature extraction has been developed for enhancing the accuracy of classification associated with place and manner of articulation information. This "elitist" approach is tested on a corpus of spontaneous Dutch using two different systems, one trained on a subset of the same corpus, the other trained on a corpus from a different language (American English). The feature dimensions, voicing and manner of articulation transfer relatively well between the two languages. However, place information transfers less well. Manner-specific training can be used to improve classification of articulatory place information.

1 Introduction

Current-generation speech recognition (ASR) systems often rely on automatic-alignment procedures to train and refine phonetic-segment models. Although these automatically generated alignments are designed to approximate the actual phones contained in an utterance, they are often erroneous in terms of their phonetic identity. For instance, over forty percent of the phonetic labels generated by state-of-the-art automatic alignment systems differ from those generated by phonetically trained human transcribers in the Switchboard corpus (Greenberg et al. 2000). The quality of automatic labeling is potentially of great significance for large-vocabulary ASR performance as word-error rate is largely dependent on the accuracy of phone recognition (Greenberg and Chang 2000). Moreover, a substantial reduction in word-error rate is, in principle, achievable when phone recognition is both extremely accurate and tuned to the phonetic composition of the recognition lexicon (McAllaster et al. 1998).

A means by which to achieve an accurate phonetic characterization of the speech signal is through the use of articulatory-acoustic features (AFs), such as voicing, place and manner of articulation, instead of phonetic segments. An advantage of using AFs is the potential performance gain for cross-linguistic transfer. Because AFs are similar across languages it should be possible, in principle, to train the acoustic models of an ASR system on articulatory-based features, independent of the language to which it is ultimately applied, thereby saving both time and effort developing applications for languages lacking a phonetically annotated set of training material.

As a preliminary means of applying AFs for cross-linguistic training in ASR, we have applied an AF-classification system originally designed for American English to spontaneous

Dutch material. This paper delineates the extent to which such cross-linguistic transfer succeeds, as well as explores the potential for applying an "elitist" approach for AF classification to Dutch. This approach improves manner-of-articulation classification through judicious (and principled) selection of frames and enhances place-of-articulation classification via a manner-specific training and testing regime.

2 Corpora

Two separate corpora, one Dutch, the other American English, were used in the study.

2.1 VIOS (Dutch)

VIOS (Strik et al. 1997) is a Dutch corpus composed of human-machine "dialogues" within the context of railroad timetable queries conducted over the telephone.

A subset of this corpus (3000 utterances, comprising ca. 60 minutes of material) was used to train an array of networks of multilayer perceptrons (MLPs), with an additional 6 minutes of data used for cross-validation purposes. Labeling and segmentation at the phonetic-segment level was performed using a special form of automatic alignment system that explicitly models pronunciation variation derived from a set of phonological rules (Kessens et al. 1999).

An eighteen-minute component of VIOS, previously hand-labeled at the phonetic-segment level by students of Language and Speech Pathology at the University of Nijmegen, was used as a test set in order to ascertain the accuracy of AF-classification performance. This test material was segmented at the phonetic-segment level using an automatic-alignment procedure, that is part of the Phicos recognition system (Steinbiss et al. 1993), trained on a subset of the VIOS corpus.

2.2 TIMIT (American English)

NTIMIT (Jankowski et al. 1990) is a quasi-phonetically balanced corpus of sentences read by native speakers of American English whose pronunciation patterns reflect a wide range of dialectal variation and which has been passed through a telephone network (i.e., 0.3-3.4 kHz bandwidth). This corpus is derived from TIMIT (an 8- kHz version of NTIMIT), which was phonetically hand-labeled and segmented at the Massachusetts Institute of Technology.

3 Training Regime

MLPs were trained on five separate feature dimensions: (1) place and (2) manner of articulation, (3) voicing, (4) rounding and (5) front-back articulation (specific to vowels), using a procedure similar to that described in (Kirchhoff 1999; Kirchhoff 2000). The front-end representation of the signal consisted of logarithmically compressed power spectra computed

over a window of 25 ms every 10 ms. The spectrum was partitioned into fourteen, 1/4-octave channels between 0.3 and 3.4 kHz. Delta (first-derivative) and double-delta (second derivative) features pertaining to the spectral contour over time were also computed. Altogether, the spectral representation was based on a 42-dimension feature space.

Articulatory-acoustic features were automatically derived from phonetic-segment labels using the mapping pattern illustrated in Table 1 for the VIOS corpus (cf. Chang et al. (2001) for the pertinent mapping pattern associated with the NTIMIT corpus). The feature dimensions, "Front-Back" and "Rounding" applied solely to vocalic segments. The approximants (i.e., glides, liquids and [h]) were classified as vocalic with respect to articulatory manner. The rhoticized segments, [r] and [R], were assigned a place feature (+rhotic) unique unto themselves in order to accommodate their articulatory variability (Lindau 1985; Vieregge and Broeders 1993). Each articulatory feature dimension also contained a class for "silence".

The context window for the MLP inputs was 9 frames (i.e., 105 ms). 200 units (distributed over a single hidden layer) were used for the MLPs trained on the voicing, rounding and front-back dimensions, while the place and manner dimensions used 300 hidden units (with a similar network architecture).

A comparable set of MLPs were trained on ca. 3 hours of material from NTIMIT, using a cross-validation set of ca. 18 minutes duration (cf. Chang et al. (2001) for additional details of this system).

Table 1: Articulatory feature characterization of the phonetic segments in the VIOS corpus. The approximants are listed twice, at top for the manner-independent features, and at bottom for manner-specific place features. The phonetic orthography is derived from SAMPA.

Consonants	**Manner**	**Place**	**Voicing**
[p]	Stop	Bilabial	-
[b]	Stop	Bilabial	+
[t]	Stop	Alveolar	-
[d]	Stop	Alveolar	+
[k]	Stop	Velar	-
[f]	Fricative	Labiodental	-
[v]	Fricative	Labiodental	+
[s]	Fricative	Alveolar	-
[z]	Fricative	Alveolar	+
[S]	Fricative	Velar	-
[x]	Fricative	Velar	+
[m]	Nasal	Bilabial	+
[n]	Nasal	Alveolar	+
[N]	Nasal	Velar	+
Approximants	**Manner**	**Place**	**Voicing**
[w]	Vocalic	Labial	+

Table 1 continued

[j]	Vocalic	High	+
[l]	Vocalic	Alveolar	+
[L]	Vocalic	Alveolar	+
[r]	Vocalic	Rhotic	+
[R]	Vocalic	Rhotic	+
[h]	Vocalic	Glottal	+
Vowels	**Front-Back**	**Place**	**Rounding**
[i]	Front	High	-
[u]	Back	High	+
[y]	Front	High	+
[I]	Front	High	-
[e:]	Front	High	-
[2:]	Front	Mid	+
[o:]	Back	Mid	+
[E]	Front	Mid	-
[O]	Back	Mid	+
[Y]	Back	Mid	-
[@]	Back	Mid	-
[Ei]	Front	Mid	-
[a:]	Front	Low	-
[A]	Back	Low	-
[Au]	Back	Low	+
[9y]	Front	Low	+
Approximants	**Front-Back**	**Place**	**Voicing**
[w]	Back	High	+
[j]	Front	High	+
[l]	Central	Mid	+
[L]	Central	Mid	+
[r]	Central	Mid	+
[R]	Central	Mid	+
[h]	Central	Mid	+

4 Cross-Linguistic Classification

Classification experiments were performed on the VIOS test material using MLPs trained on the VIOS and NTIMIT corpora, respectively (Table 2). Because ca. 40% of the test material was composed of “silence,” classification results are partitioned into two separate conditions, one in which silence was included in the evaluation of frame accuracy (+Silence), the other in which it was excluded (-Silence) from computation of frame-classification performance.

Table 2: Comparison of feature-classification performance (percent correct at frame level) for two different systems — one trained and tested on Dutch (VIOS-VIOS), the other trained on English and tested on Dutch (NTIMIT-VIOS). Two different conditions are shown — classification with silent intervals included (+Silence) and excluded (-Silence) in the test material.

	VIOS-VIOS		**NTIMIT-VIOS**	
FEATURE	**+Silence**	**-Silence**	**+Silence**	**-Silence**
Voicing	88.9	85.4	79.1	86.0
Manner	84.9	81.3	72.8	73.6
Place	75.9	64.9	52.1	38.5
Front-Back	83.0	78.0	68.9	66.9
Rounding	83.2	78.4	70.3	69.3

Classification performance of articulatory-acoustic features *trained and tested* on VIOS is more than 80% correct for all dimensions except place of articulation (cf. below for further discussion on this particular dimension). Performance is slightly higher for all feature dimensions when silence is included, a reflection of how well silence is recognized. Overall, performance is comparable to that associated with other American English (Chang et al. 2000) and German (Kirchhoff 1999) material.

Classification performance for the system trained on NTIMIT and tested on VIOS is lower than the system trained and tested on VIOS (Table 2). The decline in performance is generally ca. 10-15% for all feature dimensions, except for place, for which there is a somewhat larger decrement in classification accuracy. Voicing is the one dimension in which classification is nearly as good for a system trained on English as it is for a system trained on Dutch (particularly when silence is neglected). The manner dimension also transfers reasonably well from training on NTIMIT to VIOS. However, the place of articulation dimension does not transfer well between the two languages.

One reason for the poor transfer of place-of-articulation feature classification for a system trained on NTIMIT and tested on VIOS pertains to the amount of material on which to train. Features which transfer best from English to Dutch are those which have been trained on the greatest amount of data in English. This observation suggests that a potentially effective means of improving performance on systems trained and tested on discordant corpora would be to evenly distribute the training materials over the feature classes and dimensions classified (cf. Section 7 for further discussion on this issue).

5 An Elitist Approach to Frame Selection

With respect to feature classification, not all frames are created equal. Frames situated in the center of a phonetic segment tend to be classified more accurately than those close to the segmental borders (Chang et al. 2000; Chang et al. 2001). This “centrist” bias in feature classification is paralleled by a concomitant rise in the “confidence” with which MLPs

Table 3: The effect (in percent correct) of using an elitist frame-selection approach on manner classification for two different systems — one trained and tested on Dutch (VIOS), the other trained on English (NTIMIT) and tested on Dutch (VIOS). "All" refers to using all frames of the signal, while "Best" refers to the frames exceeding the 0.7 threshold.

	Trained and Tested on Dutch									
	Vocalic		**Nasal**		**Stop**		**Fricative**		**Silence**	
	All	**Best**	**All**	**Best**	**All**	**Best**	**All**	**Best**	**All**	**Best**
Vocalic	**89**	**94**	04	03	02	01	03	02	02	01
Nasal	15	11	**75**	**84**	03	02	01	00	06	03
Stop	16	12	05	03	**63**	**72**	07	06	10	07
Fricative	13	09	01	00	02	01	**77**	**85**	07	04
Silence	04	02	02	01	02	01	02	01	**90**	**94**

	Trained on English, but Tested on Dutch									
	Vocalic		**Nasal**		**Stop**		**Fricative**		**Silence**	
	All	**Best**	**All**	**Best**	**All**	**Best**	**All**	**Best**	**All**	**Best**
Vocalic	**88**	**93**	03	02	05	03	03	02	00	00
Nasal	46	48	**48**	**50**	02	01	02	01	01	01
Stop	22	24	10	08	**45**	**46**	21	20	02	02
Fricative	21	19	01	00	07	04	**70**	**77**	00	00
Silence	07	05	04	02	08	05	09	06	**72**	**81**

classify AFs, particularly those associated with manner of articulation. For this reason the output level of a network can be used as an objective metric with which to select frames most "worthy" of manner designation.

The efficacy of frame selection for manner classification is illustrated in the top half of Table 3 for a system trained and tested on VIOS. By establishing a network-output threshold of 0.7 for frame selection, it is possible to improve the accuracy of manner classification between 5 and 10%, thus achieving an accuracy level of 84 to 94% correct for all manner classes except stop consonants. The overall accuracy of manner classification increases from 85% to 91% across frames. Approximately 15% of the frames fall below threshold and are discarded from further consideration (representing 5.6% of the phone segments).

The bottom half of Table 3 illustrates the frame-selection method for a system trained on NTIMIT and tested on VIOS. The overall accuracy at the frame level increases from 73% to 81% using the elitist approach (with ca. 19% of the frames discarded). However, classification performance does not appreciably improve for either the stop or nasal manner classes.

6 Manner-Specific Articulatory Place Classification

Place-of-articulation information is of critical importance for classifying phonetic segments correctly (Greenberg and Chang 2000; Kirchhoff 1999) and therefore may be of utility in enhancing the performance of automatic speech recognition systems. In the classification experiments described in Section 4 and Table 2, place information was correctly classified for only 65-76% of the frames associated with a system trained and tested on Dutch. Place classification was even poorer for the system trained on English material (39-52%). A potential problem with place classification is the heterogeneous nature of the articulatory-acoustic features involved. The place features for vocalic segments (in this study, they are low, mid, and high) are quite different than those pertaining to consonantal segments such as stops (labial, alveolar, velar). Moreover, even among consonants, there is a lack of concordance in place of articulation (e.g., the most forward constriction for fricatives in both Dutch and English is posterior to that of the most anterior constriction for stops).

Such factors suggest that articulatory place information is likely to be classified with greater precision if performed for each manner class separately (cf. (Chang et al. 2001)). Figure 1 illustrates the results of such manner-specific, place classification for a system trained and tested on Dutch (VIOS). In order to characterize the *potential* efficacy of the method, manner information for the test material was derived from the reference labels for each segment rather than from automatic classification.

Five separate MLPs were trained to classify place-of-articulation features — one each for the consonantal manner classes of stop, nasal and fricative — and two for the vocalic segments (front-back and height). The place dimension for each manner class was partitioned into three features. For consonantal segments the partitioning corresponded to the *relative* location of maximal constriction — anterior, central and posterior. For example, the bilabial feature is the most anterior class for stops, while the labio-dental feature corresponds to the anterior feature for fricatives. In this fashion it is possible to construct a relational place-of-articulation customized to each consonantal manner class. For vocalic segments, front vowels were classified as anterior and back vowels as posterior. The height dimension is orthogonal to the front-back dimension and corresponds to the traditional concept of vowel height (most closely associated with the frequency of the first formant).

Figure 1 illustrates the gain in place classification performance (averaged across all manner classes) when the networks are trained using the manner-specific scheme. Accuracy increases between 10 and 20% for all place features, except “low” (where the gain is 5%).

Assigning the place features for the “approximants” (liquids, glides and [h]) in a manner commensurate with vowels (cf. Table 1) results in a dramatic increase in the classification of these features (Figure 2), suggesting that this particular manner class may be more closely associated with vocalic than with consonantal segments.

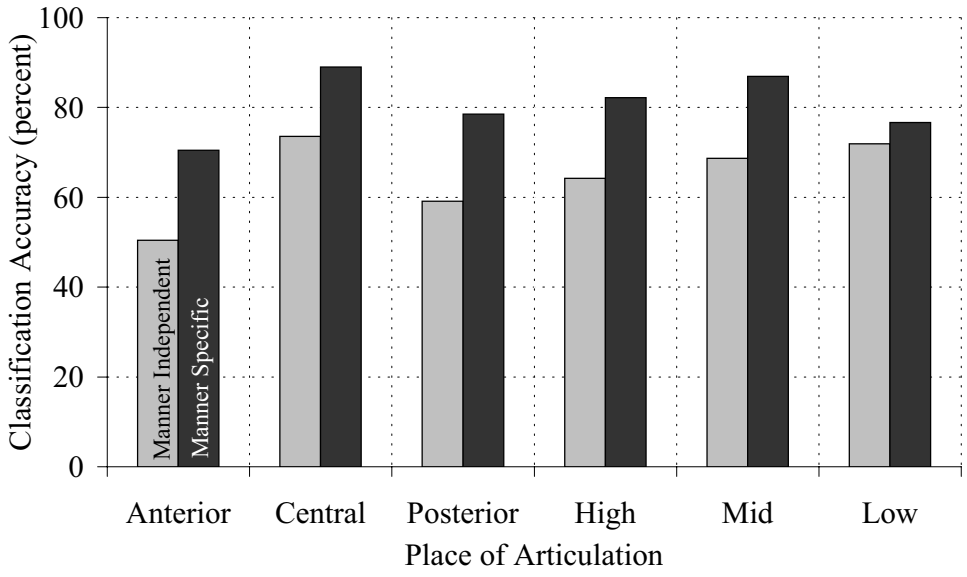


Figure 1: Comparison of place-of-articulation classification performance for two different training regimes, one using conventional, manner-independent place features (grey), the other using manner-specific (black) place features as described in Section 6. The feature classification system was trained and tested on the VIOS corpus.

7 Discussion and Conclusions

Articulatory-acoustic features provide a potentially efficient means for developing cross-linguistic speech recognition systems. The present study demonstrates that certain AF dimensions, such as voicing and manner of articulation, transfer relatively well between English and Dutch. However, a critical dimension, place of articulation, transfers much less well. An appreciable enhancement of place-of-articulation classification results from manner-specific training, suggesting that this method may provide an effective means of training ASR systems of the future.

Several challenges remain to be solved prior to deploying manner-specific, place-trained classification systems. Currently, for a (relatively small) proportion of phonetic segments (6%) the elitist approach discards all frames, thus making it difficult to recover place information for certain segments of potential importance.

A second challenge relates to the dependence of the method on the amount of training material available. AFs associated with large amounts of data usually are classified much more accurately than features with much less training material. Some means of compensating for imbalances in training data is essential.

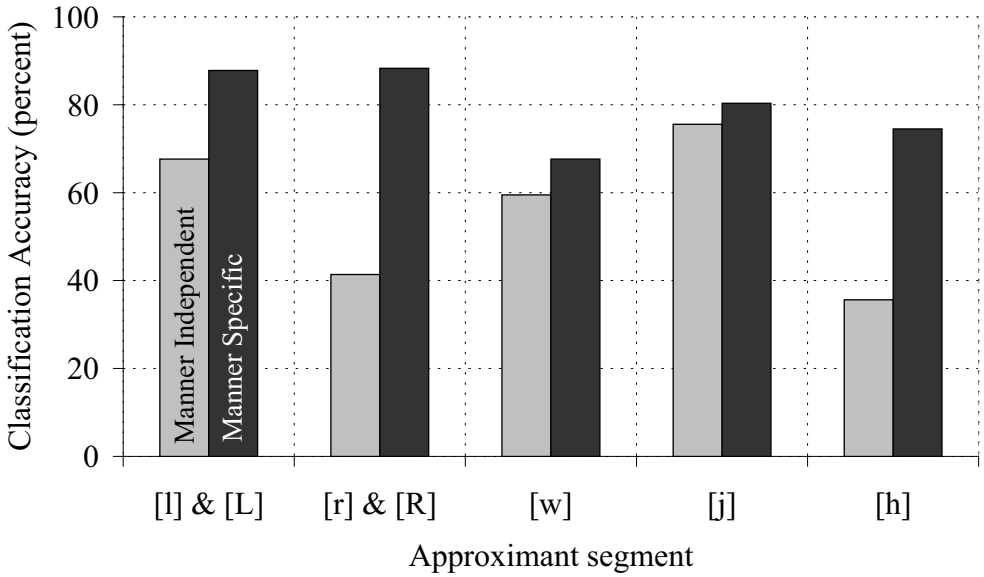


Figure 2: Comparison of manner-independent (grey) and manner-specific (black) place-trained features for the approximant subset of VIOS segments.

Finally, some means of utilizing AFs for speech recognition needs to be developed beyond the current method of merely mapping articulatory features at the frame level to the appropriate phonetic segment. Although the elitist approach provides a significant improvement of AF classification accuracy, linear mapping of the resulting AFs to phonetic segments increases phonetic-segment classification by only a small amount, (from 65% to 68%) suggesting that phonetic segments should not be the sole unit used for automatic speech recognition.

8 Acknowledgements

The research described in this study was supported by the U.S. Department of Defense and the National Science Foundation. The first author is affiliated with A^2RT, Department of Language and Speech, Nijmegen University.

References

Chang, S., S. Greenberg, and M. Wester (2001). An elitist approach to articulatory-acoustic feature classification. In *Proc. of EUROSPEECH '01*, Aalborg, pp. 1729–

1733.

Chang, S., L. Shastri, and S. Greenberg (2000). Automatic phonetic transcription of spontaneous speech (American English). In *Proc. of ICSLP '00*, Volume IV, Beijing, pp. 330–333.

Greenberg, S. and S. Chang (2000). Linguistic dissection of Switchboard-corpus automatic speech recognition systems. In *Proc. of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium*, Paris, pp. 195–202.

Greenberg, S., S. Chang, and J. Hollenback (2000). An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems. In *Proc. of the NIST Speech Transcription Workshop*, College Park, MD.

Jankowski, C., A. Kalyanswamy, S. Basson, and J. Spitz (1990). NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Proc. of ICASSP '90*, pp. 109–112.

Kessens, J., M. Wester, and H. Strik (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication 29*, 193–207.

Kirchhoff, K. (1999). *Robust Speech Recognition Using Articulatory Information*. Ph. D. thesis, University of Bielefeld.

Kirchhoff, K. (2000). Integrating articulatory features into acoustic models for speech recognition. In *PHONUS 5: Proc. of the Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Institute of Phonetics, University of the Saarland, pp. 73–86.

Lindau, M. (1985). The story of /r/. In V. Fromkin (Ed.), *Phonetic Linguistics: Essays in honor of Peter Ladefoged*, pp. 157–168. Orlando, Fl.: Academic Press.

McAllaster, D., L. Gillick, F. Scattone, and M. Newman (1998). Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In *Proc. of ICSLP '98*, Sydney, pp. 1847–1850.

Steinbiss, V., H. Ney, R. Haeb-Umbach, B.-H. Tran, U. Essen, R. Kneser, M. Oerder, H.-G. Meier, X. Aubert, C. Dugast, and D. Geller (1993). The Philips research system for large-vocabulary continuous-speech recognition. In *Proc. of EUROSPEECH '93*, Berlin, pp. 2125–2128.

Strik, H., A. Russel, H. van den Heuvel, C. Cucchiarini, and L. Boves (1997). A spoken dialogue system for the Dutch public transport information service. *International Journal of Speech Technology 2*(2), 119–129.

Vieregge, W. H. and T. Broeders (1993). Intra- and interspeaker variation of /r/ in Dutch. In *Proc. of EUROSPEECH '93*, Berlin, pp. 267–270.

Samenvatting (Summary in Dutch)

In dit proefschrift wordt onderzoek beschreven naar het modelleren van uitspraakvariatie ten behoeve van automatische spraakherkenning van het Nederlands. Het doel van automatische spraakherkenning (ASH) is om op basis van het akoestisch signaal te bepalen welke woorden, die ieder voor zich opgebouwd zijn uit een rij klanken, een spreker heeft uitgesproken. Iedere keer dat een woord geuit wordt kan de uitspraak anders zijn; dit noemen we uitspraakvariatie. De aanwezigheid van uitspraakvariatie kan tot fouten in de herkenning leiden. Het doel van dit onderzoek was om de prestatie van Nederlandse ASH te verbeteren door middel van het modelleren van uitspraakvariatie; d.w.z. het aantal correct herkende woorden binnen een van te voren vastgestelde test set te vergroten. Het type uitspraakvariatie waarop we ons in dit onderzoek hebben gericht is uitspraakvariatie die beschreven kan worden als inserties, deleties en substituties van fonen ten opzichte van een kanonieke (normatieve) transcriptie.

Spraakherkenning bestaat uit twee fases: training en herkenning. Tijdens de trainingsfase bouwt het systeem de kennis op die nodig is om spraak te herkennen. Een grote hoeveelheid opgenomen spraakmateriaal is nodig om de herkenner te trainen. In dit proefschrift is het spraakmateriaal dat we gebruikt hebben afkomstig van OVIS (Openbaar Vervoer Informatie Systeem). Het spraakmateriaal is voorzien van een orthografische transcriptie (woordelijke neerslag van hetgeen er gezegd is). Naast de orthografische transcriptie is er ook een meer gedetailleerde representatie van het materiaal nodig op het niveau van de spraakklanken. De basiseenheden die we gebruiken om de spraak te beschrijven zijn fonen, m.a.w. iedere spraakklank wordt door een foonsymbool beschreven. In het lexicon staat voor ieder woord de orthografische transcriptie met de bijbehorende foontranscriptie. Tijdens de trainingsfase wordt voor ieder woord de foontranscriptie in het lexicon opgezocht. Het trainingsmateriaal wordt vervolgens automatisch gesegmenteerd op foonniveau met behulp van het Viterbi algoritme en op basis van deze segmentatie wordt voor iedere foon een akoestisch model getraind. Dit proces wordt iteratief uitgevoerd. In de eerste stap wordt een lineaire segmentatie opgeleverd. Vervolgens worden de akoestische modellen (die getraind zijn op basis van de vorige segmentatie) gebruikt om een nieuwe segmentatie te genereren die weer gebruikt wordt om nieuwe akoestische modellen te trainen. Dit iteratieve proces gaat door tot er convergentie plaatsvindt, d.w.z. dat de segmentatie niet veel meer verandert. In totaal worden

er 39 akoestische modellen getraind: 37 foonmodellen, een model om stilte te modelleren en een model voor ruis. Daarnaast wordt er ook een taalmodel getraind dat bestaat uit een unigram (kans op een woord) en een bigram (kans op een sequentie van twee woorden). Na het voltooien van de training bestaat de spraakherkenner uit de akoestische modellen, het taalmodel en het lexicon. In de herkenningsfase wordt geprobeerd een onbekende uiting te herkennen, middels de drie onderdelen.

Een voorbeeld dat verduidelijkt waarom uitspraakvariatie kan leiden tot fouten in de herkenning is het volgende (zie ook Fig. 1.3). Stel een spreker heeft het woord “latere” uitgesproken als /lA:tr@/. In het kanonieke lexicon[1] staat alleen de fonetische transcriptie /la:t@r@/ voor het woord “latere”. Deze transcriptie in het lexicon komt niet overeen met de uitspraak van de spreker. Het gevolg hiervan kan zijn dat een onjuist woord herkend wordt, omdat er in het lexicon een andere transcriptie aanwezig is die nauwkeuriger aansluit bij het akoestisch signaal; bijvoorbeeld de transcriptie /la:tst@/ voor het woord “laatste”.

Tijdens de trainingsfase kan uitspraakvariatie leiden tot vervuilde akoestische modellen als de kanonieke transcriptie gebruikt wordt als uitgangspunt. Stel dat het zojuist genoemde voorbeeld zich in het trainingsmateriaal voordoet in plaats van in het testmateriaal. Tijdens training zou de discrepantie tussen wat er is uitgesproken en de transcriptie tot gevolg hebben dat het akoestisch model voor /@/ vervuild raakt. Delen van het spraaksignaal waarin /t/ en /r/ uitgesproken zijn worden dan gebruikt om het akoestisch model voor /@/ te trainen. Deze vervuiling van de akoestische modellen kan tot herkenfouten leiden.

Het doel van het modeleren van uitspraakvariatie is het verkleinen van het aantal fouten dat door het ASH systeem gemaakt wordt. In dit onderzoek proberen we het aantal herkenfouten te verminderen door de discrepantie tussen het akoestisch signaal en de corresponderende fonetische transcriptie te minimaliseren.

Naast een inleidend hoofstuk, dat hierboven kort is samengevat, bestaat het proefschrift uit een viertal publicaties. De eerste en derde publicatie gaat over onderzoek dat tot doel had uitspraakvariatie te modelleren, m.a.w. het minimaliseren van de discrepantie tussen het akoestisch signaal en de bijbehorende foontranscripties. In de tweede publicatie is een studie beschreven waarin de prestaties van geforceerde herkenning zijn onderzocht. Geforceerde herkenning vormt een cruciaal onderdeel van de methode waarmee wij uitspraakvariatie modeleren. De laatste publicatie gaat over het gebruik van articulatorisch-akoestische kenmerken in ASH. Met articulatorisch-akoestische kenmerken worden kenmerken als stemhebbendheid, plaats en manier van articulatie etc. bedoeld. Hieronder worden korte samenvattingen van de publicaties gegeven, gevolgd door de algemene conclusies van dit proefschrift.

Artikel 1: Een kennisgebaseerde methode voor het modelleren van uitspraakvariatie in het Nederlands.
In dit artikel is beschreven hoe de prestaties van een Nederlandse continue spraakherkenner (CSH) zijn verbeterd door het modeleren van uitspraakvariatie. In het kort bestaat de methode uit het toevoegen van varianten aan het lexicon, het hertrainen van de foonmodellen en het gebruik van taalmodellen waaraan uitspraakvarianten toegevoegd zijn.

[1] Het kanonieke lexicon bevat één transcriptie per woord.

Twee typen uitspraakvariatie zijn gemodelleerd: binnenwoord variatie en tussenwoord variatie (d.w.z. variatie die plaatsvindt over woordgrenzen heen). Binnenwoord uitspraakvarianten zijn gegenereerd door een set van vijf optionele fonologische regels toe te passen op de woorden in het kanonieke lexicon. De vijf regels zijn: /n/-deletie, /t/-deletie, /r/-deletie, /@/-deletie en /@/-insertie. Tussenwoord variatie is op twee verschillende manieren gemodelleerd. Allereerst door uitspraakvariatie over woordgrenzen als een bijzondere type binnenwoord variatie te behandelen, en ten tweede door het toevoegen van multiwoorden en hun bijbehorende varianten. Een multiwoord is een concatenatie van een reeks frequent voorkomende woorden tot één nieuw woord. De tussenwoord variatie is variatie die het gevolg is van processen zoals clitizatie, reductie en samentrekkingen.

Mogelijke uitspraakvarianten zijn verkregen door toepassing van de fonologische regels en processen op de woorden in het kanonieke lexicon. Vervolgens zijn de nieuwe transcripties aan het lexicon toegevoegd. Om uitspraakvariatie in de foonmodellen en in het taalmodel te kunnen modelleren moet eerst een geforceerde herkenning uitgevoerd worden op het trainingsmateriaal. Voor geforceerde herkenning van het trainingsmateriaal is de orthografische transcriptie van de uitingen nodig. De woorden die herkend kunnen worden tijdens een geforceerde herkenning zijn beperkt tot alleen die woorden die in de uiting voorkomen. Omdat de orthografie al bekend is, wordt de herkenner als het ware geforceerd om tussen de verschillende uitspraakvarianten van de woorden in de uiting te kiezen. Dit levert een transcriptie op die nauwkeuriger is dan een kanonieke woordtranscriptie. De nieuwe transcriptie wordt gebruikt voor het hertrainen van de foonmodellen. Op basis van de nieuwe transcriptie kan de frequentie van de uitspraakvarianten vastgesteld worden en kunnen de uitspraakvarianten met hun waarschijnlijkheden aan het taalmodel worden toegevoegd.

De sets binnenwoord en tussenwoord varianten zijn in isolatie getest maar ook in combinatie. Het foutenpercentage (op woord niveau) voor de uitgangspositie[2] was 12,8%. Voor de testconditie waarin de variatie op alle drie de niveaus in de herkenner gemodelleerd was, werd een kleine maar statistisch significante verbetering van 0,7% gemeten ten gevolge van het modelleren van alleen binnenwoord variatie. Voor de tussenwoord variatie werden kleine, statistisch niet significante verbeteringen gevonden. Het combineren van de binnenwoord varianten met de tussenwoord varianten (multiwoorden aanpak) leverde het beste resultaat op. Er werd een absolute verbetering van 1,1% ten opzichte van de uitgangspositie gemeten. Dit komt overeen met een relatieve verbetering van 8,8%.

Artikel 2: Het verkrijgen van fonetische transcripties: een vergelijking tussen expert luisteraars en een continue spraakherkenner (CSH)

In dit artikel hebben we specifiek gekeken naar het gebruik van geforceerde herkenning voor het verkrijgen van fonetische transcripties. Twee experimenten zijn uitgevoerd waarin de prestaties van een CSH vergeleken zijn met de prestaties van ervaren luisteraars. De transcriptietaak voor de luisteraars en de CSH bestond uit het aangegeven of een specifiek foon wel of niet gerealiseerd was in een uiting.

In het eerste experiment voerden de CSH en negen ervaren luisteraars dezelfde taak

[2] De uitgangspositie is de testconditie waarin geen uitspraakvariatie is gemodelleerd.

uit: ze moesten beslissen of een foon (een /n/, /r/, /t/ of /@/) wel of niet aanwezig was in 467 gevallen. Een aantal vergelijkingen tussen de oordelen van de CSH en die van de luisteraars zijn uitgevoerd. Allereerst is de overeenstemming tussen CSH-luisteraar paren vergeleken met luisteraar-luisteraar paren. De resultaten van deze vergelijkingen lieten zien dat er significante verschillen tussen de CSH en de luisteraars bestaan maar ook dat er tussen verschillende luisteraars significante verschillen bestaan. Op basis van de oordelen van de negen luisteraars was het mogelijk om referentietranscripties vast te stellen die gebaseerd waren op het meerderheidsoordeel van de luisteraars. De overeenstemming tussen de referentietranscriptie en de CSH neemt toe naarmate de referentietranscriptie strenger is, d.w.z naarmate er meer luisteraars het met elkaar eens zijn. Verder is ook de overeenstemming per fonologische regel bepaald. De vergelijkingen tussen de CSH en de luisteraars per regel lieten zien dat er voor /r/-deletie en schwa-insertie geen significante verschillen tussen luisteraars en CSH waren. Voor de andere drie processen waren de verschillen wel significant. Verder is gebleken dat de luisteraars over het algemeen meer inserties en minder deleties detecteerden dan de CSH.

In het tweede experiment is het eerste experiment verder uitgewerkt. Twee van de vijf fonologische processen zijn nader bekeken: /@/-deletie en /@/-insertie. Dit experiment is uitgevoerd om te achterhalen waarom en op welke manier de detectie van een foon door de CSH verschilt van detectie door de luisteraars. Om dit experiment uit te kunnen voeren was een meer gedetailleerde transcriptie nodig. Om deze reden hebben we een consensustranscriptie gebruikt in plaats van een transcriptie die gebaseerd is op het meerderheidsoordeel van de luisteraars. De resultaten van het tweede experiment wezen uit dat de CSH en de luisteraars verschillende drempels hebben voor het detecteren van een foon.

Op basis van de resultaten van deze experimenten concluderen we dat de geforceerde herkenning kan worden gebruikt om automatisch fonetische transcripties te verkrijgen. Ondanks het feit dat er significante verschillen tussen de CSH en de luisteraars bestaan, kunnen de verschillen acceptabel zijn, afhankelijk van het doel waarvoor de transcripties nodig zijn. De verschillen die gevonden zijn tussen de CSH en de luisteraars worden voor een deel ook tussen verschillende luisteraars gevonden.

Artikel 3: Uitspraakvariatie modellering voor ASH - kennisgebaseerde en datagestuurde methodes

In dit artikel hebben we twee verschillende methodes voor het modelleren van uitspraakvariatie bestudeerd: een kennisgebaseerde en een datagestuurde. Deze methodes verschillen in de manier waarop de informatie over de uitspraakvariatie verkregen wordt. De kennisgebaseerde aanpak bestaat in ons geval uit het gebruik van fonologische regels voor het genereren van uitspraakvarianten. De datagestuurde methode bestaat uit het uitvoeren van een vrije foonherkenning gevolgd door het gebruik van beslisbomen om varianten te generen. De twee methodes voor het modelleren van uitspraakvariatie zijn met elkaar vergeleken.

Het gebruik van kennisgebaseerde modellering had een kleine verbetering in de foutenpercentages tot gevolg. Iets grotere verbeteringen werden gevonden door het gebruik van de datagestuurde methode. Naast het vergelijken van foutenpercentages hebben we

geanalyseerd in welke mate dezelfde uitspraakvariatie wordt gemodelleerd door deze twee methodes. Het bleek dat 10% van de varianten die met behulp van de fonologische regels zijn gegenereerd ook gevonden worden in de uitvoer van de vrije foonherkenning. Dit percentage neemt toe tot 28% als beslisbomen gebruikt worden om varianten te genereren. Dit toont aan dat de beslisbomen kunnen generaliseren en dat zij varianten genereren die in het trainingsmateriaal niet geobserveerd zijn. Dit is een voordeel t.o.v. alleen gebruik te maken van vrije foonherkenning waarbij niet geobserveerde varianten niet aan het lexicon toegevoegd kunnen worden.

In dit artikel is ook een verwarbaarheidsmaat geïntroduceerd die gebruikt wordt om de verwarbaarheid binnen een lexicon te bepalen en om verwarbare varianten uit een lexicon te verwijderen. Het toepassen van deze verwarbaarheidsmaat resulteerde in ongeveer dezelfde foutenpercentages als de methode waarbij beslisbomen gebruikt werden om varianten te genereren.

Tenslotte is er een vergelijking gemaakt tussen twee verschillende typen herkenners, met het doel vast te stellen of de datagestuurde methode daadwerkelijk uitspraakvariatie modelleert of dat deze methode slechts de idiosyncratische eigenschappen van de herkenner in kwestie modelleert. De twee verschillende systemen zijn de ICSI herkenner, een hybride systeem dat gebruik maakt van neurale netten en HMMs en de Phicos herkenner, een puur HMM-gebaseerd systeem. Er zijn geen significante verschillen gevonden tussen de resultaten die met de twee verschillende herkenners gevonden zijn. Er kan dus geconcludeerd worden dat met deze datagestuurde aanpak ook daadwerkelijk uitspraakvariatie gemodelleerd wordt.

Artikel 4: Een toepassing van de “Elitist” methode voor het classificeren van articulatorisch-akoestische kenmerken van Nederlandse data.

In dit artikel is allereerst onderzocht of neurale netten die getraind zijn voor het classificeren van articulatorisch-akoestische kenmerken van Engelse data ook gebruikt kunnen worden om Nederlandse data te classificeren.

Voor zowel Nederlandse als Engelse data zijn neurale netten getraind voor de volgende vijf dimensies: (1) plaats en (2) manier van articulatie, (3) stemhebbendheid, (4) ronding en (5) voor-achter articulatie. De kenmerken ‘ronding’ en ‘voor-achter’ hebben alleen betrekking op vocalen. De articulatorisch-akoestische kenmerken zijn direct afgeleid van de foontranscripties. Bijvoorbeeld de foon /b/ zou de volgende labels krijgen: (1) bilabiaal, (2) plosief, (3) +stem, (4) n.v.t., (5) n.v.t.

Meer dan 80% van de Nederlandse data (op frameniveau) werd door een voor het Nederlands getraind systeem voor alle dimensies correct geclassificeerd, behalve voor de dimensie ‘plaats van articulatie’. Als een neuraal net getraind op Engelse data voor de classificatie van de Nederlandse data gebruikt wordt, blijken de dimensies ‘stem’ en ‘manier van articulatie’ redelijk goed overdraagbaar te zijn van het Engels naar het Nederlands, terwijl opnieuw ‘plaats van articulatie’ erg slecht geclassificeerd wordt.

Verder is in dit artikel onderzocht hoe goed de “elitist” methode werkt voor het classificeren van articulatorisch-akoestische kenmerken voor het Nederlands. Deze aanpak verschilt

van andere methodes voor het classificeren van articulatorische-akoestische kenmerken doordat er manier-specifieke training van plaats van articulatie wordt gedaan. Twee belangrijke observaties liggen ten grondslag aan deze aanpak. Allereerst, de observatie dat frames die zich in het midden van een fonetisch segment bevinden vaker correct en met een hogere waarschijnlijkheid geclassificeerd worden dan de frames die zich dicht bij de segmentgrenzen bevinden. Dit blijkt vooral te gelden voor 'manier van articulatie'. Ten tweede, 'plaats van articulatie' wordt erg slecht geclassificeerd. Een belangrijke reden hiervoor is de heterogene aard van deze dimensie. Deze twee observaties hebben geleid tot de manier-specifieke training van plaatskenmerken. De resultaten die in dit artikel gepresenteerd zijn wijzen uit dat in principe substantiële verbeteringen in de classificatie van 'plaats van articulatie' haalbaar zijn met deze aanpak.

Algemene conclusies

Het in dit proefschrift beschreven onderzoek laat zien dat methodes die voor Engelse ASH ontwikkeld zijn ook voor het Nederlands toepasbaar zijn. Het hoofddoel van dit onderzoek was om de herkenprestaties van een Nederlands ASH systeem te verbeteren door het modeleren van uitspraakvariatie. Statistisch significante verbeteringen zijn gevonden door middel van kennisgebaseerde en datagestuurde modelleermethodes.

Een andere vraag die onderzocht is in dit proefschrift is of de geforceerde herkenning die in dit uitspraakvariatieonderzoek gebruikt is, ook zinvol toegepast zou kunnen worden om fonetische transcripties te verkrijgen voor linguïstisch onderzoek. Een vergelijking tussen de transcripties die verkregen zijn door geforceerde herkenning en transcripties verkregen door luisteraars, laat zien dat er significante verschillen zijn tussen de transcripties van de herkenner en die van de luisteraars, maar ook dat er significante verschillen tussen de luisteraars onderling bestaan. Ondanks deze (significante) verschillen kunnen fonetische transcripties verkregen met geforceerde herkenning acceptabel zijn, afhankelijk van het doel waarvoor de transcripties nodig zijn.

Een beperking van geforceerde herkenning voor het verkrijgen van fonetische transcripties is dat er een orthografische transcriptie voor nodig is. Het gebruik van articulatorisch-akoestische kenmerken zou dit probleem kunnen omzeilen. In de laatste publicatie hebben we laten zien dat het in principe mogelijk is om nauwkeurige transcripties te genereren zonder gebruik te maken van een orthografische transcriptie. Deze methode moet echter nog verder ontwikkeld en verfijnd worden om uiteindelijk tot volledige transcriptie op woordniveau te komen.

Curriculum Vitae

Mirjam Wester was born August 24^{th}, 1971 in Delft, the Netherlands. She attended Aiyura International Primary School in Papua New Guinea. For her secondary education, she attended three different schools; first Ukarumpa High School in Papua New Guinea, which was followed by three years at the Gertrudis Lyceum, Roosendaal, the Netherlands, and in 1990 she graduated from the Schothorst College, Amersfoort, the Netherlands. She went on to study Phonetics at the University of Utrecht, the Netherlands. From August 1992 to January 1993 she was enrolled at the University of Stockholm, Sweden in the framework of the Erasmus exchange program. Her Master's thesis involved research into automatic classification of voice quality using regression models and hidden Markov models. She passed her *Doctoraal Examen* in Phonetics (equivalent to obtaining an MA degree) in 1996, having specialized in Language and Speech Processing. From January 1997 to June 2001 Mirjam Wester was employed at the Department of Language and Speech at the University of Nijmegen, the Netherlands, as a PhD student (AIO). During this period she spent a year at the International Computer Science Institute (ICSI) in Berkeley, California, as a visiting researcher. This thesis is a result of the work carried out at the Department of Language and Speech at the University of Nijmegen and at ICSI. Mirjam Wester is currently employed as a researcher in the MUMIS project (Multimedia Indexing and Searching Environment) at the University of Nijmegen's Department of Language and Speech.